In Silico Design and Selection of CD44 Antagonists

Ruiz Moreno, Angel

DOI:
10.33612/diss.193912062

In Silico Design and Selection of CD44 Antagonists

Implementation of computational methodologies in drug discovery and design

Angel Jonathan Ruiz Moreno

2021

university of groningen

The research presented in PhD this thesis was performed at Laboratorio de Farmacología Molecular (Molecular Pharmacology Lab), School of Medicine, Universidad Nacional Autónoma de México (UNAM) and at Drug Design Lab within the Groningen Research Institute of Pharmacy (GRIP), University of Groningen (RUG), The Netherlands under the conditions of a double degree PhD program agreed between UNAM and RUG (52392-1452-8-V-18). The research work was carried out according to the requirements of the double degree PhD agreement.

The projects were supported at UNAM by: PAPIIT UNAM IN219719, PAPIIT UNAM IV200121, CONACYT A1-S-18285, LANCAD-UNAM-DGTIC-364 2018–2019, and LANCAD-UNAM-DGTIC-386 2020-2021.

The projects were supported at RUG by: ITN "AcceleratedEarly stage drug dIScovery" (AEGIS, grant agreement No. 675555), COFUND ALERT (grant agreement No. 665250), KWF Kankerbestrijding grant (grant agreement No. 10504), National Institute of Health (NIH) (2R01GM097082-05), European Lead Factory (IMI) (grant agreement number 115489), Qatar National Research Foundation (NPRP6-065-3-012), Prominent (grant agreement no. 754425), Marie Skłodowska-Curie (grant agreement no. 675555), and Hartstichting (ESCAPE-HF, 2018B012)

Angel J. Ruiz-Moreno received the graduate scholarship 584534 from CONACYT, funding from *Programa de Apoyo a los Estudios de Posgrado* (PAEP), UNAM 2018-2019 and fundings from the Maria Sybilla Merian (MSM) budget from Faculty of Science and Engineering (FSE) or RUG to support his PhD studies.

Printing of this thesis was financially supported by the University Library and the Graduate School od Science, Faculty of Science and Engineering, University of Groningen, The Netherlands.

Cover Design: Angel Jonathan Ruiz Moreno

university of groningen

In Silico Design and Selection of CD44 Antagonists

Implementation of computational methodologies in drug discovery and design

PhD thesis

to obtain the degree of PhD at the
University of Groningen
on the authority of the
Rector Magnificus Prof. C. Wijmenga
and in accordance with
the decision by the College of Deans.

This thesis will be defended in public on

Friday 10 December 2021 at 09:00 hours

by

Angel Jonathan Ruiz Moreno

born on 1 April 1990 in
Mexico City, Mexico.

Supervisors

Prof. A.S.S. Dömling
Prof. M.A. Velasco-Velázquez

Assessment committee

Prof. C. Camacho
Prof. S.J. Marrink
Prof. G.J. Poelarends

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

DOCTORADO EN CIENCIAS BIOMÉDICAS

FACULTAD DE MEDICINA

DISEÑO Y SELECCIÓN IN SILICO DE ANTAGONISTAS DE CD44

(IN SILICO DESIGN AND SELECTION OF CD44 ANTAGONISTS)

TESIS
QUE PARA OPTAR POR EL GRADO DE:
DOCTOR EN CIENCIAS

PRESENTA
ANGEL JONATHAN RUIZ MORENO

DIRECTOR DE TESIS
DR. MARCO A. VELASCO VELÁZQUEZ, FACULTAD DE MEDICINA, UNAM.
PROF. DR. ALEXANDER DÖMLING, DRUG DESIGN, UG.

COMITÉ TUTOR (UNAM)
DR. ALFREDO TORRES LARIOS
INSTITUTO DE FISIOLOGÍA CELULAR
DRA. LETICIA ROCHA ZAVALETA
INSTITUTO DE INVESTIGACIONES BIOMÉDICAS

CIUDAD DE MÉXICO, MÉXICO. DICIEMBRE DE 2021

Comité evaluador:

Presidente: Dra. Nuria Victoria Sánchez Puig
Secretario: Dr. Marco Antonio Velasco Velázquez
Vocal: Dr. Martin González Andrade
Vocal: Dr. Matthew Groves
Vocal: Dr. Enrique Ramon Ángeles Anguiano

All members of assessment committee (Comité evaluador) at UNAM and RUG revised this same thesis prior its submission to the libraries.

To my Family and Friends.

To everyone who is still standing next to me, as well as those who have left.

...las flores como los humanos, compartimos el mismo destino…

Content of the thesis

Chapter 1

(Introductory chapter)

Implementation of computational methodologies for drug discovery

Angel J. Ruiz-Moreno, Marco A. Velasco-Velázquez, and Alexander Dömling.

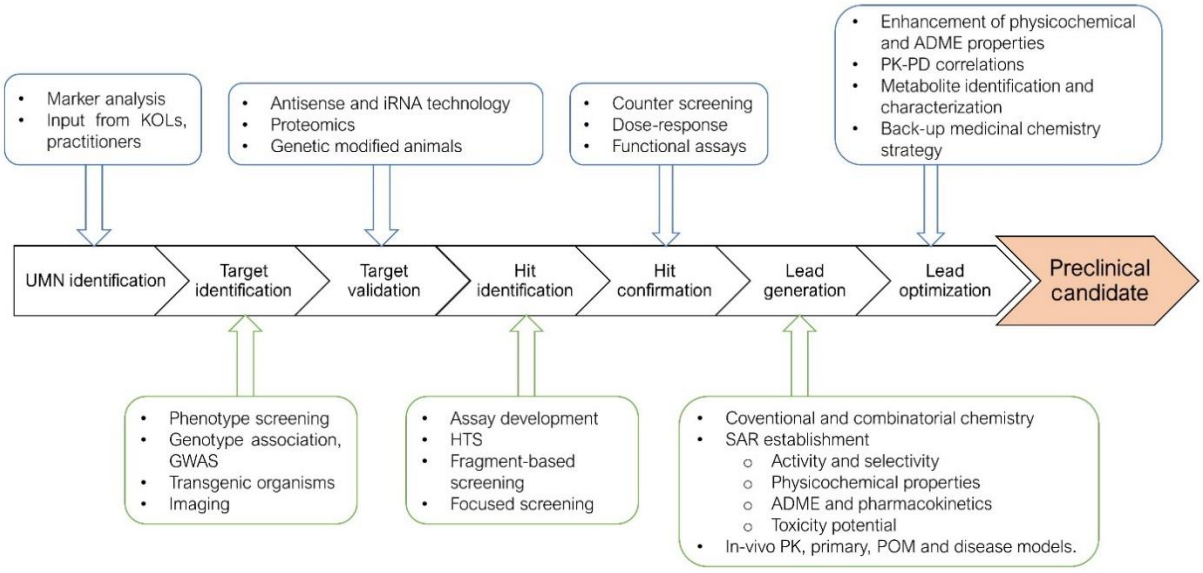
Part of this chapter is being prepared for publication

Abstract

Drug discovery is a process that aims to identify drug candidates by a thorough evaluation of the biological activity of small synthetic molecules or biomolecules. The modern drug discovery process includes identifying the disease to be treated and its unmet medical needs; selecting a druggable molecular target and validating it; developing *in vitro* assays for high throughput screening of compound libraries to identify hits against the target; hit optimization to generate lead compounds with adequate efficacy, potency, and selectivity in animal models. Subsequently, the lead compounds are further optimized to improve their potency and pharmacokinetics before moving forward with clinical development. Computational strategies are now necessary tools for speeding up multiple steps of the drug discovery process. The use of such approaches is described in this chapter.

Introduction

Drug discovery is a process that aims to identify a small synthetic molecule or a large biomolecule drug candidate after a thorough evaluation. In general, the preclinical drug discovery process includes multiples steps (Figure 1): identifying the disease to be treated and its unmet medical needs; selecting a druggable molecular target and validating it; developing *in vitro* assays for high throughput screening of compound libraries against the target to identify that alter target´s activity (called hits); hit optimization to generate lead compounds with adequate *in vitro* potency and selectivity, and demonstrate efficacy and safety in disease animal models. Subsequently, the lead compounds are further optimized to improve their efficacy and pharmacokinetics before moving forward to clinical drug development [1].



Overview of preclinical drug discovery process. UMN, unmet medical needs; KOL, key opinion leader; SAR, structure-activity relationship; GWAS, genome-wide association studies; HTS, high throughput screening; POM, proof of mechanism; PK, pharmacokinetics; PD, pharmacodynamics (modified from [1]).

The drug´s target

Drugs typically fail in the clinic for two main reasons; primarily is that they lack efficacy, and the second is that they are not safe [1,2]. As such, identification and validation of the target is one of the most crucial steps in developing a new drug. “Target” is a general term that refers to a variety of biological entities, such as proteins, genes, and RNA [3]. A good target must be efficient in modulating the disease, meet clinical and commercial requirements and, most importantly, be ‘druggable.’ A ‘druggable' target can be reached by the putative drug molecule, which is usually a small molecule or a larger biological one, and which, once bound, elicits a biological response that can be measured both *in vitro* and *in vivo* [2,3].

Target identification is the process of selecting the molecular target - for example, a protein or a nucleic acid - of a small molecule [4]. Introducing a small molecule drug into a living cell is the same as directly disrupting one or more nodes in a complex network of genes and gene products. In general, this perturbation causes a response in almost every node of the network (all genes and their products). The task of target identification is to identify direct target nodes and distinguish them from nodes that respond indirectly to the target nodes. The traditional approach to target identification is to employ a variety of experimental methods to locate the active compound's physical binding sites [5]. However, physical binding determination, in addition to being a costly and time-consuming process that is often the rate-limiting step in drug development, does not always imply an observable effect on protein or gene activity. As a result, it is preferable to identify targets based on actual cell response data [5]. This prompted the proposal of model-based methods in which response data are used to infer network models, including the direct targets of external perturbations, for the underlying gene regulatory networks [4–6].

Drug target validation could be highly beneficial for new drug research and development and for gaining a better understanding of the pathogenesis of the target-related diseases. Essentially, the target validation process should include identifying a biomolecule of interest. To assess its potential as a target, a bioassay to measure relevant biological activity must be developed. Then, it is possible to conduct a high-throughput screening (HTS) to find hits. Hits can be further characterized using the same or additional bioassays [7–9].

The target validation process should be carried out on three levels: molecular, cellular, and whole animal model levels. Small molecules obtained through HTS are useful for confirming new drug targets. The majority of HTS models are molecular in nature, i.e., cell-free systems. Screening for a specific enzyme inhibitor, for example, usually entails mixing the enzyme with samples to detect a decrease in the substrate or an increase in the product in the enzyme catalytic process. However, there is a significant difference between a cell-free system and a cell-based system. Because there are many predictable and unpredictable factors, the results from this level are not entirely reliable. Actual results from this level, on the other hand, convey the point that hits genuinely act on the target. Validation at the cellular level confirms cell-free results. Small molecules may be used at this level to highlight the pathological significance of the target. Animal

models are used to validate the target at the whole level. If the hit obtained from HTS displays a therapeutic effect in animal models, then it may be promising [9,10]. Target identification and validation improves confidence in the target-disease relationship and allows researchers to investigate whether target modulation causes mechanism-based side effects [2,3].

Computer-Aided Drug Discovery

Computer-aided drug discovery (CADD) is a discipline that comprises a broad range of theoretical and computational approaches that are part of modern drug discovery. CADD collects multiple chemical-molecular and quantum strategies to discover, design, and developing therapeutic chemical agents. Many CADD approaches are based on structure-activity relationships (SAR). The main objectives of CADD are part of multidisciplinary work for the improvement of bioactive molecules, the development of therapeutic alternatives, and the understanding of biological events at the molecular level [11].

The CADD methods can be broadly classified into two groups, namely structure-based (SB) and ligand-based (LB) drug discovery. CADD has become an effective and indispensable tool in therapeutic development. The human genome project has made a large amount of sequence data available for use in various drug discovery projects. Furthermore, increased knowledge of biological structures and increased computer power have enabled the effective use of computational methods in various stages of the drug discovery and development pipeline. As a result, *in silico* tools are more important than ever before, and they have advanced pharmaceutical research. As a result, CADD has been integral in discovering numerous available pharmaceutical drugs that have received FDA approval and have reached the consumer market [12–14].

In the case of proteins as drug targets, especially those for which their three-dimensional (3D) structure has been solved, the SB methods are fundamental. Although structures of relevant drug targets were not always available directly from X-ray crystallography in the early days, comparative models based on homologs began to be used in lead optimization in the 1980s [15,16]. It was recognized in the 1990s that 3D structures were helpful in defining topographies of the complementary surfaces of ligands and their protein targets and that they could be used to optimize the potency and selectivity of the leads [17]. As a result, several crystal structures of real drug targets became available. For example, the anti-AIDS drugs Agenerase and Viracept were

developed using the crystal structure of HIV protease [18,19]. The use of crystallographic structures has also been employed to develop treatments for other diseases such as influenza [20] and cancer [21,22]. Several drugs are currently on the market as a result of this structure-based design approach [23].

SB designing methods, which were previously used to optimize these leads into drugs, are now frequently used much earlier in the drug discovery process. Protein structure is used in target identification and selection (the assessment of a target's druggability), virtual screening for hits, and fragment screening. Furthermore, structural biology's critical role in lead optimization to engineering increased affinity and selectivity into leads remains as crucial as ever [16].

This chapter provides an overview of the structure-based methods used in CADD. We discuss structure prediction tools that are commonly used in structure-based drug discovery, molecular docking algorithms, virtual screening as a high-throughput screening technique, lead optimization, methods for evaluating drug ADME properties, and the use of molecular dynamics (MD) for the study of protein-ligand interactions and binding affinity prediction. In addition, we will discuss recent advancements and implementation of Artificial Intelligence (AI) algorithms for CADD.

Predicting the target structure by sequence/structure homology

The function of a protein is primarily determined by its 3D structure. However, methods for determining the spatial organization of a protein are time-consuming and expensive. The structure determination process typically entails the development of a protein expression system, protein purification crystallization, and finally, structure determination, which each successive step may take months to years to accomplish [24]. For this reason, while the number of available protein sequences has increased exponentially, the number of experimentally derived protein structures has lagged far behind. There has been extensive research into *in silico* methods for structure determination over several decades. The sequence/structure homology approach aims to create a method for determining the 3D structure of a protein based solely on its sequence [25]. One strategy, known as homology modeling, takes advantage of protein structure redundancy by predicting the structure of an unknown protein using homologous proteins, or structurally related proteins from the same family. Even though there are millions of proteins, the number of distinct structural folds is two to three orders of magnitude lower [25,26]. The assumption is that all

members of a protein family are related through divergent evolution from a common ancestor and thus must share the same basic fold [27]. Thus, if a protein belongs to a family in which the experimental structures of several proteins have been determined, an atomic model of the protein can be constructed by comparing it to those structures. The structural genomics initiatives aim to characterize most protein sequences through an efficient combination of targeted high-throughput experimental structure determination and prediction [28], implying that homology modeling is a useful tool for biologists [25].

SWISS-MODEL (https://swissmodel.expasy.org/) is a well-known and widely used online tool for protein structure homology modeling due to its extensive functionality, ease of use, and speed. The prediction procedure includes template recognition, target-template alignment, model construction, and model evaluation. For template recognition and target-template alignment, BLAST [29] and HHblits [30] are employed. The target protein structure is constructed by copying the atomic coordinates from the template following target-template alignment. The unaligned region is constructed by searching the fragment library. The quality of the final model is evaluated by the knowledge-based score function QMEAN [31]. The broad functionality, easy operability, and fast running make SWISS-MODEL one of the most widely-used homology modeling tools in the past two decades [32–34].

Another homologous modeling method is *Modeller* (https://salilab.org/modeller/), which implements structure modeling by satisfying spatial constraints. The spatial constraints can be derived from target-template alignment as well as data from NMR spectroscopy, fluorescence spectroscopy, site-directed mutagenesis, stereochemistry, and intuition. Stereochemical restraints usually supplement these homology-derived restraints on bond lengths, bond angles, dihedral angles, and nonbonded atom-atom contacts that are obtained from molecular mechanics force field. The final model is evaluated by DOPE potential [35]. Modeler can also perform auxiliary tasks such as fold assignment, phylogenetic tree calculation, and *de novo* modeling of loops in protein structures [36,37].

The use of homology models in drug discovery has played an important role in developing kinase inhibitors [38]. Over 500 kinases have been identified in the human genome, but only 50 have had their structures determined experimentally so far. When the structure of an inhibitor is unknown, homology models can be used to optimize its affinity [38]. This includes increasing

affinity for the kinase target(s) and decreasing affinity for related kinases and identifying kinases for selectivity screens based on structural homology in the region where the inhibitor binds [39]. Homology models can also be successfully used as a starting point for virtual screening.

Identifying ligand-binding regions

several efforts have been made to generate structural data of protein targets using experimental or computational models. It is becoming increasingly important to develop functional computational methods capable of identifying sites involved in forming intermolecular interactions in protein-ligand systems [16]. For instance, by identifying steric strain or other types of high-energy conformations that frequently occur at active sites [40,41] or by identifying protein clefts that can accommodate ligands [42]. Almost all protein functional sites are typically formed through mutation and Darwinian selection, and thus they are the most highly conserved regions of a protein [43–45]. Several computational methods have been developed to integrate this information to predict putative binding sites in proteins [46].

The realistic prediction of putative ligand or protein binding sites has important implications for rational drug design and a better understanding of protein-protein interaction networks. Identifying and characterizing putative protein binding sites on proteins can assist in determining the number and type of protein interaction partners. *In silico* methods for predicting protein-protein interaction sites can also be used to predict protein aggregation [47,48] or the nonspecific binding to a variety of partners [49]. Recent approaches to predicting protein binding sites and protein binding sites and which partner protein may bind could be useful in predicting protein interaction networks [50].

High-affinity binding cavities for small drug-like ligands, contrary to protein-protein sites, are frequently less polar or more hydrophobic than the rest of the protein surface [42,47,51–55]. Because organic drug-like molecules are smaller in size than proteins, the buried surface area of small molecule protein-ligand interactions is generally smaller than in protein-protein interactions. Small molecules' high-affinity binding sites are usually concave pockets or cavities on the surface of proteins, or sometimes partially buried that achieve strong interactions through a sufficiently large number of favorable protein-ligand contacts [51]. In many ways, algorithms for predicting protein-protein interfaces are similar to methods for predicting binding regions for small drug-like

molecules. However, due to the distinct general architecture of these types of binding sites, there are some significant differences [54,56].

Since a concave binding cleft on the protein surface commonly forms the small-molecule binding site, the detection of binding pockets or protein cavities requires special attention. Generally, the importance of a concave binding site explains the better performance of binding site prediction for small drug-like ligands compared to protein binding site prediction. Concave regions on protein surfaces are less common than the flat or slightly curved surfaces typical for protein-protein interfaces [46]. The explanation and methods of several algorithms based on various detection principles for druggable binding site detection have been previously reviewed in detail [57].

Fpocket (http://fpocket.sourceforge.net/) is an open-source pocket detection package that uses Voronoi tessellation and alpha spheres to detect pockets. The modular tool is organized around a central library of functions, which serves as the foundation for three main programs: i) Fpocket is used to identify protein pockets, (ii) Tpocket is used to organize pocket detection benchmarking on a set of known protein-ligand complexes, and (iii) Dpocket is used to collect pocket descriptor values on a set of proteins. Fpocket relying on a simple scoring function can detect 94% and 92% of the pockets within the best three ranked pockets from the holo and apoproteins, respectively. Fpocket provides a rapid, open-source, and stable basis for further developments related to protein pocket detection, efficient pocket descriptor extraction, or druggability prediction purposes [58].

Structure-guided virtual screening

Multiple techniques are considered to examine the interplay between alteration in protein biological functionality and therapeutic consequences. As a result, druggability uses a structure-based approach to assess the likelihood that small drug-like molecules have intrinsic potencies to bind and modulate protein activities. [59–61].

The available public domain databases that are specifically aimed at drug discovery scientists all have their own specialist content and, in general, this is complementary. For example, vendor information, patented compounds, data on marketed drugs, and bioactivity data for both efficacy and liability targets, and crystal structures of small molecules bound to protein targets can all be

found in public databases. The number of compounds ranges from the comparatively small manually curated sets, such as ChEBI, to large patent databases, such as SureChEMBL, where the data is extracted from patents using 'name to structure' and 'image to structure' software and for which manual curation would be a prohibitively expensive task. Some databases also take depositions from other databases or directly from depositors. For example PubChem includes data from ChEMBL and ChEBI, alongside an extensive set of user depositions; ChEMBL includes some data from PubChem, and ZINC contains data from ChEMBL [63].

When selecting compounds in the early stages of the drug discovery phase, drug-likeness is essential because it helps optimize pharmacokinetics and pharmaceutical attributes as chemical stability, solubility, bioavailability, and distribution profile [62]. Drug-like compounds are chemical molecules with functional groups and physical properties that are similar to the majority of known drugs and thus can be hypothesized to be biologically active or have therapeutic potential [64,65]. Compared to the narrow range distribution of approved drugs, these compounds fall below essential physicochemical thresholds such as molecular mass, hydrophobicity, and polarity [66,67]. Regardless, there is no apparent structural similarity between drug-like compounds. Any approved drug and drug development has a high rate of attrition; the selective selection of compounds with inherent drug-likeness improves the chances of surviving this event [67]. A thorough understanding of the principles of target binding site recognition and druggability, as well as the drug-likeness of a chemical molecule, could thus increase the likelihood of novel chemical molecules becoming viable therapeutic options in disease intervention. [61].

Selecting (or designing) compounds *in silico* that bind to a protein active or druggable site can represent a challenge [16]. Structure-guided virtual screening (VS) is a complementary tool to HTS that attempts to find hits in the early stages of drug discovery. Specifically, once a macromolecular target is selected, compounds are needed to initiate efforts toward a clinical candidate. The goal of VS is to identify these early "hits" among a library of compounds. What differentiates HTS from VS is that HTS is an experimental approach, while VS is a theoretical one. HTS tests large numbers of compounds for their ability to affect the activity of target molecules by addressing whether a compound reacts biochemically with the target. For example, questions such as "does it bind to the target protein?", "does it trigger enzymatic reactions?", "does it activate signaling pathways?" are explored by HTS assays. As mentioned above, compounds showing positive results are passed onto a more rigorous assay. It cannot be emphasized enough that positive results must be reconfirmed because if false positives are being pursued, the

investment detriment down the road will be high. On the other hand, negative results can mean that a potentially valuable compound is not considered. The latter could be an issue if no hits are found. However, the goal of HTS is not to find all possible hits in a library collection but a sufficiently enough set to use as starting scaffolds for initial discovery efforts. Therefore, while false positives could be costly if pursued and thus, re-confirmation of the results is critical, false negatives are not and should not cause worry. In the following step, chemists need to intelligently select two to three classes of compounds that show the most promise for potential clinical candidates. HTS is time-consuming, requires infrastructure, and has a low success rate (<5%); nonetheless, it has been the method of choice for the last 20 years in the pharmaceutical sector.

On the other hand, VS is an *in-silico* HTS method. It consists of virtually placing (docking) collections of millions of compounds into a biological target, followed by an evaluation of the tightness of the fit (scoring). VS offers a quick assessment of huge libraries and reduces the number of compounds that need testing to identify early hits. Thus, the basic requirements for VS are: i) a compound collection, which highly depends on the objective of the project (see above); ii) the structure of the biological target, and iii) an appropriate docking/scoring scheme. The choices made for each of these requirements come with a set of questions that need to be addressed to make the process as efficient and accurate as possible [68].

A great variety of docking tools and programs, such as *AutoDock* [69], *AutoDock Vina* [70], *LeDock* [71], *UCSF DOCK* [72], *Glide* [73], *GOLD* [74], *MOE Dock* [75], *Surflex-Dock* [76], etc., have been developed for both commercial and academic uses [77–79]. The sampling algorithm and scoring function, which determine a docking program's sampling and scoring power, are the two most important components. Shape matching, systematic search (such as exhaustive search, fragmentation, and conformational ensemble), and stochastic search algorithms(such as Monte Carlo methods, genetic algorithms, Tabu search methods, and swarm optimization methods) are the most commonly used sampling algorithms [80]. Popular scoring functions can be divided into three categories: force field, empirical, and knowledge-based scoring functions [81–83]. Recently, some quantum mechanical (QM) and semi-empirical quantum mechanical (SQM) scoring functions have been developed to capture binding affinity trends and native pose identification [84,85]. The problem of sampling efficiency can be effectively, or at least partially, overcome with the rapid development of computer hardware. However, it remains a huge challenge for available scoring functions to predict the binding affinities of diverse small molecules with high accuracy [86,87].

When tested on large sets of protein-ligand complexes, state-of-the-art docking programs correctly dock 70–80% of ligands [73,74,88]. However, in docking techniques, it is complex to determine a ligand's relative affinity compared to other compounds. In most virtual screening approaches, *in silico* methods dock and rank many commercially available (or synthetically accessible) compounds, and the highest-ranking compounds are chosen for acquisition (or synthesis) and experimental testing for action against the target protein. Despite these difficulties, the *in silico* methods are valuable and influential in structure-guided design. For example, if there is an enormous enrichment of actual hits in a selected subset of compounds than a subset selected by another mechanism, VS would be effective. This condition does not necessitate a highly accurate scoring feature or a rigorous treatment of receptor flexibility. Successful examples of virtual screening in identifying novel hits and the demonstration of significant enrichment have been described, and there have been many other reports in the literature [68,89–91].

It is interesting to consider current and future trends in the VS approach. There is still much to be done in the area of improved ranking and scoring functions. However, while this method is beneficial, it is not completely satisfactory [92–94]. Receptor flexibility is still a hot topic of study, and the current standard is to use functions that tolerate minor clashes and multiple receptor conformations [95–97]. Water molecules that form hydrogen bonds with proteins are now being studied using automated methods that enable ligands to either displace or form hydrogen bonds with them, depending on which case is the most energetically favorable [98]. Given the academic and pharmaceutical industry's success with structure-based drug design and virtual screening, several more examples of virtual screening and docking applications can be expected to appear in the literature in the future [16].

De novo structure-guided drug design

Existing compound libraries cannot always contain molecules that are an optimal fit for a given target protein, or the compounds themselves may be a limited novelty. *In silico* methods that can design new ligands are, therefore, potentially useful. One of the *de novo* design methods involves placing fragments in the binding cleft of protein targets and then ‘growing' them to fill the available space, optimizing electrostatic, van der Waals, and hydrogen bonding interactions [99–102]. There are significant challenges in correctly positioning and scoring modeled molecules, just as there are with virtual screening. However, the fragment-based *de novo* design can also generate

molecules that are difficult to synthesize [101,102]. For these reasons, there has been a trend to start *de novo* design with a moiety (or fragment) that has previously been shown to bind well to the protein target. Since the designed compounds are built on an active scaffold, more control over the 'synthesizability' of the designed compounds can be exercised. As a result, the designed compounds are more likely to bind to the receptor.

The evaluation of candidate compounds is critical in the design process because a de novo design program typically suggests many candidate compounds. Scoring functions indicate which structures are the most promising. In this context, it is common to distinguish between receptor-based (e.g., docking, receptor-derived pharmacophores) and ligand-based (e.g., similarity metrics) scoring, depending on the reference knowledge used to guide the search for new compounds [103]. In addition, multiple scoring functions can be used in parallel to enable multi-objective design [104,105], i.e., different (potentially competing) properties considered simultaneously. For example, properties like aqueous solubility, toxic characteristics, synthetical feasibility, and biological activity are of vital interest for potential ligand structures and they can be explicitly incorporated into the construction process by multi-objective scoring [106].

In the ideal case, *de novo* design software suggests high-quality (in terms of the scoring function) molecular structures. However, there is no guarantee that a designed compound will receive immediate approval from a medicinal chemist. It is critical to understand that *de novo* design rarely results in new chemotypes with nanomolar activity, target selectivity, and an acceptable pharmacokinetic profile. Instead, de novo generated molecules are frequently "concept compounds" that require significant further optimization. However, compared to screening an arbitrary compound collection, de novo design can be expected to have a higher hit rate. A *de novo* design algorithm must solve three problems for successful automated compound design [107]:

i) The structure sampling problem – how to assemble candidate compounds, e.g., atom-based or fragment-based.

ii) The scoring problem – how to assess molecule quality, for example, through 3D receptor-ligand docking and scoring (requires receptor structure) or ligand-based similarity measure (requires reference ligands, also known as "templates").

iii) The optimization problem – how to navigate in search space systematically, for example, using depth-first/breadth-first search, Monte Carlo sampling with the Metropolis criterion, evolutionary algorithms, or exhaustive structure enumeration.

The majority of the early *de novo* design tools were entirely atom-based. Modern approaches frequently include a wide range of large and small virtual molecular entities for compound construction, as well as a few single-atom fragments. Atom-based approaches have the advantage of performing fine-grained molecule sculpting and, theoretically, assembling the entire chemical universe of structures. These benefits come at a cost: many potential solutions complicate a systematic search for beneficial compounds. The fragment-based approach, which reduces the search space size significantly, is a shortcut to generating new ligands. If fragments that are commonly found in drug molecules are used for molecule assembly, the designed compounds have a high likelihood of being druglike themselves. Notably, fragment hits have a high "ligand efficiency" [108,109] (calculated as binding energy divided by the number of non-hydrogen atoms), making them ideal for further optimization. A fragment can be as simple as a single atom or as complex as a polycyclic ring system. It is insufficient to have a high probability of exhibiting desired biological activity on the target for *de novo* designed candidate ligands.

As previously mentioned, proposed compounds must also be chemically synthesizable. Since the early days of de novo design, it has been well established that one of the primary goals of de novo design is the ease of synthesis of the virtually constructed molecules. Nonetheless, for a long time, this issue has gone unaddressed. The chemical feasibility of candidate compounds is still a problem that is far from being solved. The assembly of molecular building blocks by rule-based virtual reaction schemes is a typical pattern of *de novo* design programs that explicitly consider synthetic tractability. Suitable building blocks, for example, can be obtained through the virtual retrosynthesis of drug molecules. The same set of reactions are then used to assemble new candidate compounds. It is reasonable to expect such designed compounds to have some degree of "drug-likeness" and to contain only a few undesirable (e.g., reactive, toxic) structural elements [110]. Virtual structure assembly can be guided by simulated organic synthesis steps, allowing a synthesis route for each generated structure to be proposed. Alternatively, additional software can automatically analyze de novo design ligand candidates to propose generalized synthetic routes and select potential reagents from databases of available compounds [106].

A strategy that combines the use of relevant scaffolds and fragments to obtain compounds within a limited chemical space is computational combinatorial chemistry (CCC). Over the last two decades, the number of synthesis protocols and chemical scaffolds available for parallel, automated, fast chemistry has dramatically increased [111,112]. Particularly useful are reactions that embrace a large number of starting materials while remaining resistant to structural variations within their respective reactive groups, enabling the resulting compound library to cover a wide range of structural diversity. For example, libraries can be created with theoretical multicomponent reactions (MCRs). In MCRs, the combination of three or more different starting material groups allows for the quick and easy creation of large combinatorial libraries that accept a wide range of chemical functionality. MCR libraries have been shown to yield drug candidates from various chemical scaffolds [112–114]. Combinatorial library screening (either virtual or experimental) has proved to be an effective method in the hit-and-lead discovery phase. The success rate of designed biased and focused compound libraries can reach over 50%, significantly higher than the historical rate of 15–20% for “blind” high-throughput screening [112]. It turns out that the use of structural information corresponds to a notable change in combinatorial chemistry in recent years, away from broad "diversity" library synthesis and toward far more focused approaches.

Molecular dynamics in drug discovery

The introduction of computers and efficient computational science tools have played a critical role in speeding up the drug discovery process while also making it less cumbersome and less expensive. An active area of research is the development of efficient methods and computational tools to aid in the speeding up of quantitative *in silico* calculations of drug and target properties beyond binding prediction. For example, many quantum mechanical and classical simulation tools are now available to assist in calculating the properties of molecules that can be used to predict their potency as a drug [115].

In molecular dynamics (MD) experiments, a computer simulation is used to gain insight into the behavior of a real-world physical system or process. A model system is created to accomplish that specific goal that represents or emulates the given physical system [116]. The target structure is not fixed but fluctuates in response to temperature. Thus, using a single static target structure, such as a crystallographic structure, to predict binding sites and protein-ligand binding from VS experiments may result in an inaccurate prediction of binding sites and protein-ligand

binding [115]. Accordingly, for the model system, a MD simulation and analysis should generate a time series or an ensemble of states ("observations") and provide various system properties derived from these states [116].

More importantly, using MD, one can observe behavior that is otherwise inaccessible to experimentation and test “what-if” scenarios (e.g., mutation studies) [117,118]. As a result, simulation techniques have become invaluable tools for modern research, complementing experimental approaches. With the continued advance of computing power, such tools will only become more important [116]. MD seeks to derive statements about a molecular system's structural, dynamical, and thermodynamic properties. The system is typically a biomolecule (solute) immersed in an aqueous solvent, such as a protein, enzyme, or a collection of lipids forming a membrane (water or electrolyte). For proteins and enzymes, the experimental protein structure deposited in the Protein Data Bank (PDB) [119] serves as a starting point for MD simulations. If no structure is available, one must resort to modeling (predicting) the structure for which several techniques are available and have previously been described in this text. MD simulations could be used to generate a set of target conformations and assess the ability of ligands to bind to target proteins. A traditional MD trajectory will be made up of successive target conformations that are very close to the local minimum corresponding to the starting conformation.

However, several MD methodologies have been developed to simulate other relevant effects such as binding energy calculation or ligand-driven conformational changes of the target, such methods include replica exchange MD [120], accelerated MD [121], and metadynamics [122]. MD simulations can be used to accurately predict a drug molecule's binding site and corresponding binding energy.

MD simulation techniques have established their relevance in modern drug discovery and development processes over the last 5 decades [115,116]. It is now possible to simulate even large and complex (bio)molecular systems in time scales that can provide important insights into real-life molecular events and interactions, thanks to advances in computer power, new force fields, and improved sampling methods. Successful MD of the protein RAS revealed the conformational dynamics of highly flexible target proteins, providing useful information for drug discovery and development [123–125]. Furthermore, MD-generated conformational ensembles

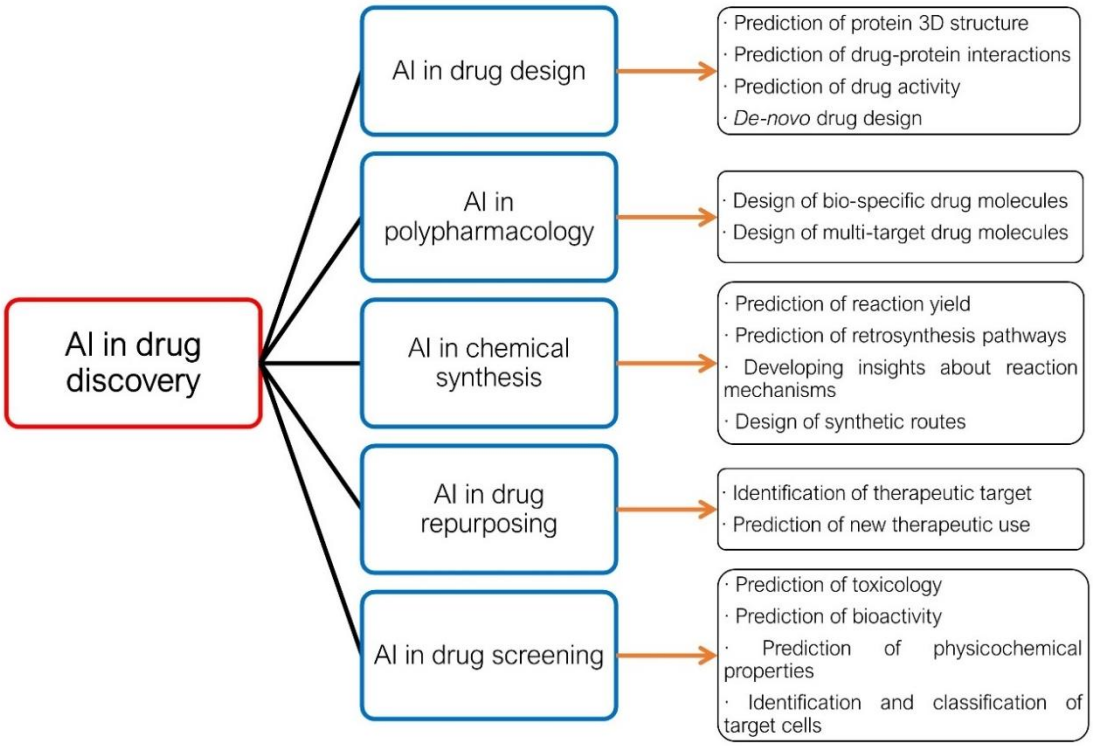
can aid ligand design against intrinsically disordered proteins that are extremely flexible [126–128]. MD simulations can also reveal important details about antibody-antigen interactions, which can aid in the development of new antibody therapeutics with better properties [129–131]. To improve the accuracy of ranking putative hits, MD-based binding free energy calculations are widely used in the hit identification phase [132–134].

Furthermore, various enhanced MD techniques that probe the drug residence times at the target site aid in developing drugs with improved binding kinetics [135]. In developing peptide therapeutics, MD simulations can also help with peptide docking by improving conformational sampling and refinement of peptide-protein complexes [136,137]. Due to the large size of the simulation system, MD simulations of membrane proteins embedded in a relevant cell membrane model have been complicated, but those using a method called coarse-graining have helped to reduce the computational cost [138,139]. The addition of the membrane to MD simulations revealed new information about lipid-protein interactions, membrane protein function, and ligand entry and exit into the target binding site [140,141]. MD simulations have also been shown to be useful in pharmaceutical development and formulation studies, in addition to the first steps of the drug discovery process. For example, crystalline and amorphous drugs [142], drug-polymer formulations [143], or drug-loaded nanoparticles [144] can be studied by MD simulations to complement experimental studies. Insights into such systems at the molecular level can improve the solubility, stability, and other properties of drug formulations [116].

In summary, MD simulations can be a valuable tool for assisting in the early stages of modern drug discovery and development. Furthermore, due to theoretical and technological advancements in the field, MD simulations are likely to gain even more importance, and they play an important role in the process of drug discovery. Hopefully, one day, a computer will be able to develop a drug, significantly speeding up the process of drug discovery [115,116].

Artificial Intelligence in drug discovery

The vast chemical space, which is estimated to contain over $>10^{60}$ molecules, encourages the testing of many molecules to identify to increase the number of molecules with clinical application [145]. The lack of advanced technologies, on the other hand, limits such testing, making it a time-consuming and costly task that can be addressed by using artificial intelligence (AI) [146]. AI can recognize hit and lead compounds, allowing for faster drug target validation and structure design optimization [145,147]. Different applications of AI in drug discovery are depicted in Figure 3.



Role of artificial intelligence (AI) in drug discovery. AI can be used effectively in different parts of drug discovery, including drug design, chemical synthesis, drug screening, polypharmacology, and drug repurposing. Modified from [147].

AI is a technology-based system that uses a variety of advanced tools and networks to simulate human intelligence. AI makes use of systems and software that can interpret and learn from data in order to make independent decisions in order to achieve specific goals. Its applications are continuously being extended in the pharmaceutical field [147]. To date, rapid advances in AI-guided automation are on track to completely transform society's work culture [148,149].

AI encompasses various method domains, including reasoning, knowledge representation, solution search, and machine learning (ML). ML employs algorithms that can recognize patterns within a set of data that has been further classified. Deep learning (DL) is a subfield of machine learning that uses artificial neural networks (ANNs). These are made up of a network of interconnected sophisticated computing elements called "perceptons," which are analogous to human biological neurons and mimic the transmission of electrical impulses in the brain [150]. ANNs are a collection of nodes that receive separate inputs and convert them to outputs, either singly or in groups, using algorithms to solve problems [151]. ANNs involve various types,

including multilayer perceptron (MLP) networks, recurrent neural networks (RNNs), and convolutional neural networks (CNNs), which utilize either supervised or unsupervised training procedures [152].

AI in drug screening

Physicochemical properties of a drug, such as a solubility, partition coefficient (logP), degree of ionization, and intrinsic permeability, indirectly impact its pharmacokinetic properties and specificity for a target receptor family must be taken into account when developing a new drug [153]. Physicochemical properties can be predicted using a variety of AI-based tools. For example, ML uses large data sets generated during previous compound optimization to train the program [154]. In addition, molecular descriptors, such as SMILES strings, potential energy measurements, electron density around the molecule, and atom coordinates in 3D, are used in drug design algorithms to generate feasible molecules via DNN and predict their properties [155].

Prediction of bioactivity and toxicity

The affinity of drug molecules for the target protein or receptor determines their efficacy. Drug molecules that do not interact with or have a high affinity for the targeted protein will not be able to provide a therapeutic response. It is also possible that developed drug molecules interact with unintended proteins or receptors, resulting in potential toxicity. As a result, DTBA (drug-target binding affinity) is critical for predicting drug-target interactions. AI-based methods can assess a drug's binding affinity by looking at the features or similarities between the drug and its target. To determine the feature vectors, feature-based interactions recognize the chemical moieties of the drug and the target. On the other hand, in a similarity-based interaction, the similarity of the drug and the target is taken into account, and it is assumed that similar drugs will interact with the same targets [156].

Many strategies involving ML and DL have been used to determine DTBA. ML-based approaches such as Kronecker-regularized least squares (KronRLS) evaluate drug and protein molecule similarity [157]. Drug features from SMILES, ligand maximum common substructure (LMCS), extended connectivity fingerprint, or a combination thereof can also be considered [156]. For instance, matched molecular pair (MMP) analysis [158] investigates a single localized change to a drug candidate and its impact on the molecular properties and bioactivity of the molecule. It is

commonly used in quantitative structure-activity relationship (QSAR) studies [158]. In a typical study, MMPs for *de novo* design tasks are generated using retrosynthesis rules. A candidate molecule has a static core plus two fragments (which describe the transformation) [159]. After that, the core and these fragments are encoded. Finally, three previously used machine learning (ML) methods, namely random forest (RF) [160], gradient boosting machines (GBMs) [161], and Deep Neural Networks (DNNs) [162], are used to extrapolate to new transformations, fragments, and modifications of the static core. These models, for example, were trained using IC_{50} data from five different kinases and a bromodomain-containing protein [163]. DNN outperformed RF and GBM in terms of overall performance in predicting compound activity [163]. With the dramatic increase in public databases containing a large number of structure–activity relationship (SAR) analyses (such as ChEMBL and Pubchem), MMP with ML has been used to predict many bioactivity properties such as oral bioavailability [164], distribution coefficient (logD) [165,166], intrinsic clearance [167], absorption, distribution, metabolism, and excretion (ADME) [165,168], and mode of action [169].

AI can also be employed to predict toxic effects and, thus, to avoid undesirable characteristics of any drug molecule. In order to predict toxicity, preliminary studies using cell-based *in vitro* assays are frequently used, followed by animal studies to determine a compound's toxicity, increasing the cost of drug development. LimTox, pkCSM, admetSAR, and Toxtree are some of the AI web-based tools that can help predict toxicity, decreasing the costs during drug development [154]. Advanced AI-based approaches look for similarities among compounds or project the toxicity of the compound based on input features. The Tox21 Data Challenge was a collaboration between the National Institutes of Health (NIH), the Environmental Protection Agency (EPA), and the US Food and Drug Administration (FDA) to test several computational techniques for predicting the toxicity of 12,707 environmental compounds and drugs [154]. DeepTox, an ML algorithm, outperformed all other methods by identifying static and dynamic features within chemical descriptors of molecules, such as molecular weight (MW) and Van der Waals volume, and could accurately predict a molecule's toxicity using predefined 2,500 toxicophore features [170]. The DeepTox algorithm first normalizes the chemical representations of the compounds, from which a large number of chemical descriptors are computed and used as the input to ML methods. The descriptors are categorized as static or dynamic. Static descriptors include atom counts, surface areas, and the presence or absence of a predefined substructure in a compound [170][36]. The presence and absence of 2,500 predefined toxicophore features [38] and other chemical features extracted from standard molecular fingerprint descriptors are also calculated. Dynamic

descriptors are calculated in a prespecified way. Despite a potentially infinite number of different dynamic features, the algorithm keeps the dataset within manageable limits [36]. In typical test cases, the DeepTox algorithm shows good accuracy in predicting the toxicology of compounds [36].

Prediction of the target protein structure

As previously stated, it is critical to predict the target protein structure to design the drug molecule for selective disease targeting. Because the design is in accordance with the chemical environment of the target protein site, AI can assist in structure-based drug discovery by predicting the 3D protein structure, which can help predict the effect of a compound on the target as well as safety considerations prior to its synthesis or production [171]. AlphaFold, a DNN-based AI tool, was used to analyze the distance between adjacent amino acids and the corresponding angles of the peptide bonds to predict the 3D target protein structure. It performed exceptionally well, correctly predicting 25 out of 43 structures using only primary protein sequences [172]. These results were significantly better than the second-place contester, which correctly predicted only three of 43 test sequences. AlphaFold relies on DNNs that are trained to predict the properties of a protein from its primary sequence. It predicts the distances between pairs of amino acids and the φ–ψ angles between neighboring peptide bonds. These two probabilities are then combined into a score which is used to estimate the accuracy of a proposed 3D protein structure model. Using these scoring functions, AlphaFold explores the protein structure landscape to find structures that match predictions [172].

Predicting drug–protein interactions

Predicting a drug interaction with a receptor or protein is critical for understanding its efficacy and effectiveness, allowing for drug repurposing and avoiding polypharmacology [171]. Various AI methods have successfully predicted ligand-protein interactions with high accuracy, resulting in improved therapeutic efficacy [171,173]. Wang et al. used a model based on support vector machines (SVMs) to discover nine new compounds and their interactions with four key targets. The model was trained on 15,000 protein-ligand interactions and was developed based on primary protein sequences and structural characteristics of small molecules [174]. Yu et al. used two random forests (RF) models with high sensitivity and specificity to predict possible drug-protein interactions by combining pharmacological and chemical data and validating them against known platforms such as SVM. These modes could also predict drug-target associations, which

could be expanded to include target–disease, and target associations, speeding up the drug discovery process [175].

AI's ability to predict drug-target interactions has also been used to aid in repurposing existing drugs and avoiding polypharmacy. When a drug is repurposed, it automatically qualifies for Phase II clinical trials [145]. This saves money because relaunching an existing drug costs ~$8.4 million versus ~$41.3 million for launching a new drug entity. [176]. Drug–protein interactions can also help predict the likelihood of polypharmacology, which is when a drug molecule interacts with multiple receptors, resulting in off-target side effects. [177]. AI can help design safer drug molecules by designing new molecules based on the principles of polypharmacology. [178]. SOM and other AI platforms and the vast databases available can be used to link multiple compounds to a variety of targets and off-targets. Bayesian classifiers and similarity ensemble approach (SEA) algorithms can be used to establish links between the pharmacological profiles of drugs and their possible targets [179]. Li et al. showed how to use KinomeX, an AI-based online platform that uses DNNs to detect polypharmacology in kinases based on their chemical structures. This platform employs a DNN that has been trained with over ~14 000 bioactivity data points derived from over ~300 kinases. Thus, it can investigate a drug's overall selectivity for the kinase family and specific subfamilies of kinases, which can aid in developing new chemical modifiers [180]. Ligand Express, Cyclica's cloud-based proteome-screening AI platform, is one notable example. Ligand Express is used to find receptors that can interact with a specific small molecule (whose molecular description is in SMILE string) and produce on- and off-target interactions. This aids in comprehending the drug's potential side effects [181].

On the other hand, Quantum Mechanics (QM) or QM/molecular mechanics (MM) hybrid methods are useful for predicting protein-ligand (drug) interactions in drug discovery [182,183]. These methods consider quantum effects for the simulated system (or region of interest in the case of QM/MM) at the atomic level, providing significantly higher accuracy than classical MM methods. Because MM methods only use simple energy functions based on atomic coordinates, the time cost of QM-based methods is much higher than that of MM methods [183,184]. The use of AI methods in QM calculations thus involves a tradeoff between QM accuracy and the lower time cost of MM models [185]. AI models have been trained to reproduce QM energies from atomic coordinates and outperform MM methods in calculation speed. AI is mostly used for atomic simulations and electrical property predictions. In contrast, DL has been used to predict the potential energies of small molecules, effectively replacing computationally demanding quantum

chemistry calculations with a fast ML method [185]. DFT (density functional theory) potential energies derived from quantum chemistry have been calculated and used to train DNNs on large datasets. In a study of two million elpasolite crystals, for example, the accuracy of an ML model improved with increasing sample size, reaching 0.1 eV/atom for DFT formation energies trained on 10,000 structures. The model was then used to evaluate compositional alternatives for a variety of properties [186].

The de novo drug design approach has been widely used to design drug molecules in recent years. The traditional method of *de novo* drug design is being phased out in favor of evolving DL methods, which have the advantages of less complicated synthesis routes and more straightforward prediction of novel molecule bioactivity [155]. Computer-aided synthesis planning can also suggest millions of structures that can be synthesized and predict multiple synthesis routes for each of them [187].

After a molecule has been virtually screened for potential bioactivity and toxicology, the search for the best chemical synthesis pathway to synthesize the drug candidate begins. This step is frequently difficult and inefficient. Despite knowledge of hundreds of thousands of transformation steps, novel molecules cannot be efficiently synthesized due to novel structural features or conflicting reactivities [188]. Retrosynthesis analysis searches for 'backward' reaction pathways indefinitely until a set of simpler, readily available precursor molecules is obtained. [189]. Monte Carlo tree search (MCTS) [190] is the technique of choice for making branch decisions. Monte Carlo simulations employ random search steps with no branching until an optimal solution is discovered. Algorithms for computer-assisted synthesis planning (CASP) [191,192] were previously developed to aid retrosynthesis analysis, but they failed to gain widespread acceptance among chemists. These algorithms necessitate incorporating human knowledge into executable programs. However, manual encoding of chemistry does not scale to exponentially growing knowledge, and the results obtained from reaction databases frequently lacked chemical intelligence [189]. ML approaches trained on empirical data can now be used to (i) predict the likelihood of a transformation at a specific branching position and (ii) guide the selection of random steps. A predefined transformation rule can link the molecule (or an intermediate) to specific precursors at each transformation step. AI algorithms can be trained on the yields and costs of these transformation rules in the literature and then predict the most feasible retrosynthesis pathway for a given molecule.

A recently reported 3N-MCTS method [189] combines three different neural networks with MCTS to create a CASP workflow. Each network is in charge of a specific task: (i) an expansion node, (ii) a rollout node, and (iii) an update node. In the expansion node, the algorithm looks for new ways to transform the molecule (or an intermediate) in the past. It includes a 'in-scope' policy that evaluates the feasibility of a transformation using 12.4 million transformation rules from the literature [193]. The neural networks are trained to predict the best transformation for the molecule (or intermediate) in question, guiding the selection of expansion pathways. Because positive data dominate the literature, a transformation is thought to be less feasible if its reverse reaction is high yielding.

Furthermore, selecting high-yielding transformations helps to eliminate the possibility of side products [189]. In the rollout node, the 'in-scope' policy is similar to that in the expansion node, except that only frequently reported transformation rules are used. During the expansion phase, this strategy allows for a slow and thorough search for the best transformation possibilities but a faster evaluation of position values during the rollout phase. The evaluation of a specific pathway is incorporated into the search tree in the update node. These nodes are used iteratively to search for transformations with the highest scores for a molecule submitted for retrosynthesis analysis and can eventually identify possible precursors for the entire reaction pathway [189].

The performance of MCTS on the test set of molecules was superior to that of other alternative algorithms. For example, MCTS solved 80% of retrosynthesis problems when a 5 s per molecule time limit was applied [189], and the rate of solving can exceed 90% if the time limit is raised to 60 s. More impressively, the speed per molecule for 3N-MCTS is 20-fold faster than the traditional Monte Carlo method [189,194].

Grzybowski et al. developed the Chematica program [195], now renamed Synthia, which can encode a set of rules into the machine and suggest possible synthesizing routes for eight medicinally important target molecules. This program has shown to be effective in terms of increasing yield and lowering costs. It is also capable of providing alternative synthesizing strategies for patented products and assisting in the synthesis of compounds that have not been synthesized yet. Similarly, DNN focuses on organic chemistry and retrosynthesis rules, which, when combined with MCTS and symbolic AI, improves reaction prediction and the drug discovery and design process, which is much faster than traditional methods [189,196].

AI use in *de novo* molecule design can benefit the pharmaceutical industry because of its numerous advantages, including online learning and simultaneous optimization of previously learned data and suggesting possible synthesis routes for compounds, resulting in a faster lead design and development [197,198].

Conclusions

Computational methods such as protein structure prediction methods, virtual high-throughput screening, and docking methods have been used to accelerate the drug discovery process. They are now routinely used in academia and the pharmaceutical industry. These methods are well established and have shown great promise and success. They are now a valuable integral part of the drug discovery pipeline. Computationally predicting and filtering large molecular databases and selecting the most promising molecules to be optimized is less expensive and faster.

Only molecules with the predicted biological activity will be tested *in vitro*. This saves money and time by lowering the risk of committing resources to potentially ineffective compounds that would otherwise be tested in vitro.

Virtual screening methods based on structure and ligand are widely used, with most applications focusing on enzyme targets [199]. Even though structure-based methods are more commonly used, ligand-based methods have resulted in the discovery of many potent drugs. Virtual HTS (VHTS) methods are useful for quickly screening large repositories of small molecules and selecting a smaller number of potential drug-like molecules for testing. These methods can help to significantly reduce the cost associated with the drug discovery process by reducing the number of possible molecules that need to be tested experimentally. Several studies have shown that VHTS can identify molecules that conventional high-throughput screening (HTS) experiments cannot. [200]. As a result, VHTS methods are frequently used in conjunction with HTS methods. These methods select molecules that are more likely to be drug candidates and should be considered when selecting hits.

Proteins are typically represented as static structures in experimental methods. On the other hand, proteins are highly dynamic in nature, and protein dynamics play an important role in their functions. Computational modeling of proteins' flexibility is of great interest, and several ensemble-based methods in structure-based drug discovery have emerged [201]. Molecular dynamics simulations are frequently used to generate target ensembles that can then be used in molecular docking [201].

Hybrid structure-based and ligand-based methods are also gaining popularity. These combined (ligand-based and structure-based) drug discovery methods are appealing because they can enhance the benefits of both methods while also improving the protocols. [202,203]. CADD has had a significant impact on developing various therapeutics that are currently being used to treat patients. Despite its successes, CADD faces some challenges, including the accurate identification and prediction of ligand binding modes and affinities. The phenomenon of drug polymorphism is one of the most difficult aspects of drug discovery [204]. Drug polymorphism occurs when a drug has multiple forms that are chemically identical but structurally different. This can have a significant impact on the success of a drug. Solubility, stability, and dissolution rates of different polymorphic forms of a drug with different solid-state structures can differ. Drug polymorphism can have an impact on drug bioavailability, efficacy, and toxicity. If a different polymorphic form of the same drug is administered, one polymorphic form responsible for a specific drug effect may differ. It is possible to characterize drugs with different polymorphic forms using techniques such as spectroscopy.

Protein-protein interactions (PPIs) present yet another difficulty in drug discovery. PPIs play a role in a variety of cellular processes and biological functions that have been linked to disease. As a result, small molecule drugs targeting PPIs are essential in drug discovery [205]. The development of therapeutics that can either disrupt or stabilize these interactions is of interest.

However, designing inhibitors that can directly interrupt PPIs is complex. Most drugs are designed to target a specific binding site on a protein of interest. Protein–protein-interacting surfaces, on the other hand, have more extensive interfaces and are more exposed. As a result, their binding sites are frequently poorly defined. Finding the sites that can be targeted for PPI inhibition is thus difficult and crucial.

The field of CADD is constantly evolving, with advancements being made in all areas. Scoring functions, search algorithms for molecular docking and virtual screening, hit optimization, and assessment of ADME properties of potential drug candidates are among the areas of focus. With the current successes, computational methods have a promising future in assisting in discovering many more therapeutics.

Finally, the advancement of AI, along with its remarkable tools, is constantly aimed at reducing challenges faced by pharmaceutical companies, affecting the drug development process as well as the overall lifecycle of the product, which might explain the rise in the number of start-ups in this sector [194]. The current healthcare sector is confronted with many complex challenges, such as rising drug and therapy costs, and society requires significant changes in this area. Personalized medications with the desired dose, release parameters, and other required aspects can be manufactured according to individual patient need using AI in pharmaceutical product manufacturing [206]. Using the latest AI-based technologies will reduce the time it takes for products to reach the market, but it will also improve product quality and overall safety of the manufacturing process and provide better resource utilization and cost-effectiveness, highlighting the importance of automation [207].

References

1. Sinha S, Vohora D. Drug Discovery and Development. Pharmaceutical Medicine and Translational Clinical Research. Elsevier; 2018. pp. 19–32. doi:10.1016/B978-0-12-802103-3.00002-X

2. Hughes J, Rees S, Kalindjian S, Philpott K. Principles of early drug discovery. Br J Pharmacol. 2011;162: 1239–1249. doi:10.1111/j.1476-5381.2010.01127.x

3. Santos R, Ursu O, Gaulton A, Bento AP, Donadi RS, Bologa CG, et al. A comprehensive map of molecular drug targets. Nature Reviews Drug Discovery. 2017;16: 19–34. doi:10.1038/nrd.2016.230

4. Schenone M, Dančík V, Wagner BK, Clemons PA. Target identification and mechanism of action in chemical biology and drug discovery. Nature Chemical Biology. 2013;9: 232–240. doi:10.1038/nchembio.1199

5. Ursu A, Waldmann H. Hide and seek: Identification and confirmation of small molecule protein targets. Bioorg Med Chem Lett. 2015;25: 3079–3086. doi:10.1016/j.bmcl.2015.06.023

6. Jacobsen EW, Nordling TEM. Robust Target Identification for Drug Discovery. IFAC-PapersOnLine. 2016;49: 815–820. doi:10.1016/j.ifacol.2016.07.290

7. Blake RA. Target Validation in Drug Discovery. High Content Screening. New Jersey: Humana Press; 2006. pp. 367–378. doi:10.1385/1-59745-217-3:367

8. Houston JG, Banks MN. High-Throughput Screening for Lead Discovery. Burger's Medicinal Chemistry and Drug Discovery. American Cancer Society; 2003. pp. 37–69. doi:10.1002/0471266949.bmc020

9. Chen X-P, Du G-H. Target validation: A door to drug discovery. : 7.

10. Gad SC, editor. Drug Discovery Handbook: Gad/Drug Discovery Handbook. Hoboken, NJ, USA: John Wiley & Sons, Inc.; 2005. doi:10.1002/0471728780

11. Prieto-Martínez FD, López-López E, Eurídice Juárez-Mercado K, Medina-Franco JL. Computational Drug Design Methods—Current and Future Perspectives. In Silico Drug Design. Elsevier; 2019. pp. 19–44. doi:10.1016/B978-0-12-816125-8.00002-X

12. Leelananda SP, Lindert S. Computational methods in drug discovery. Beilstein J Org Chem. 2016;12: 2694–2718. doi:10.3762/bjoc.12.267

13. Clark DE. What has computer-aided molecular design ever done for drug discovery? Expert Opinion on Drug Discovery. 2006;1: 103–110. doi:10.1517/17460441.1.2.103

14. Tanaji TT, Santosh AK, Alan CR. Successful Applications of Computer Aided Drug Discovery: Moving Drugs from Concept to the Clinic. Current Topics in Medicinal Chemistry. 2009;10: 127–141.

15. Blundell TL. Structure-based drug design. Nature. 1996;384: 23–26. doi:10.1038/384023a0

16. Congreve M, Murray CW, Blundell TL. Keynote review: Structural biology and drug discovery. Drug Discovery Today. 2005;10: 895–907. doi:10.1016/S1359-6446(05)03484-7

17. Campbell SF. Science, art and drug discovery: a personal perspective. Clin Sci (Lond). 2000;99: 255–260.

18. Miller M, Schneider J, Sathyanarayana BK, Toth MV, Marshall GR, Clawson L, et al. Structure of complex of synthetic HIV-1 protease with a substrate-based inhibitor at 2.3 A resolution. Science. 1989;246: 1149–1152. doi:10.1126/science.2686029

19. Lapatto R, Blundell T, Hemmings A, Overington J, Wilderspin A, Wood S, et al. X-ray analysis of HIV-1 proteinase at 2.7 Å resolution confirms structural homology among retroviral enzymes. Nature. 1989;342: 299–302. doi:10.1038/342299a0

20. Varghese JN. Development of neuraminidase inhibitors as anti-influenza virus drugs. Drug Development Research. 1999;46: 176–196. doi:https://doi.org/10.1002/(SICI)1098-2299(199903/04)46:3/4<176::AID-DDR4>3.0.CO;2-6

21. Lombardino JG, Lowe JA. The role of the medicinal chemist in drug discovery — then and now. Nature Reviews Drug Discovery. 2004;3: 853–862. doi:10.1038/nrd1523

22. Wong S, Witte ON. The BCR-ABL story: bench to bedside and back. Annu Rev Immunol. 2004;22: 247–306. doi:10.1146/annurev.immunol.22.012703.104753

23. Hardy LW, Malikayil A. The impact of structure-guided drug design on clinical agents. 2003; 6.

24. Chayen NE, Saridakis E. Protein crystallization: from purified protein to diffraction-quality crystal. Nature Methods. 2008;5: 147–153. doi:10.1038/nmeth.f.203

25. Pitman MR, Menz RI. 2 - Methods for Protein Homology Modelling. In: Arora DK, Berka RM, Singh GB, editors. Applied Mycology and Biotechnology. Elsevier; 2006. pp. 37–59. doi:10.1016/S1874-5334(06)80005-5

26. Xu J, Li M, Lin G, Kim D, Xu Y. Protein threading by linear programming. Pac Symp Biocomput. 2003; 264–275.

27. Pandit SB, Bhadra R, Gowri V, Balaji S, Anand B, Srinivasan N. SUPFAM: A database of sequence superfamilies of protein domains. BMC Bioinformatics. 2004;5: 28. doi:10.1186/1471-2105-5-28

28. Baker EN, Arcus VL, Lott JS. Protein structure prediction and analysis as a tool for functional genomics. Appl Bioinformatics. 2003;2: S3-10.

29. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25: 3389–3402. doi:10.1093/nar/25.17.3389

30. Remmert M, Biegert A, Hauser A, Söding J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nat Methods. 2011;9: 173–175. doi:10.1038/nmeth.1818

31. Benkert P, Tosatto SCE, Schomburg D. QMEAN: A comprehensive scoring function for model quality assessment. Proteins. 2008;71: 261–277. doi:10.1002/prot.21715

32. Biasini M, Bienert S, Waterhouse A, Arnold K, Studer G, Schmidt T, et al. SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. Nucleic Acids Res. 2014;42: W252-258. doi:10.1093/nar/gku340

33. Guex N, Peitsch MC. SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. Electrophoresis. 1997;18: 2714–2723. doi:10.1002/elps.1150181505

34. Deng H, Jia Y, Zhang Y. Protein structure prediction. International journal of modern physics B. 2018;32. doi:10.1142/S021797921840009X

35. Modi V, Dunbrack RL. Assessment of refinement of template-based models in CASP11. Proteins. 2016;84 Suppl 1: 260–281. doi:10.1002/prot.25048

36. Sali A, Blundell TL. Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol. 1993;234: 779–815. doi:10.1006/jmbi.1993.1626

37. Webb B, Sali A. Comparative Protein Structure Modeling Using MODELLER. Curr Protoc Bioinformatics. 2016;54: 5.6.1-5.6.37. doi:10.1002/cpbi.3

38. Diller DJ, Li R. Kinases, homology models, and high throughput docking. J Med Chem. 2003;46: 4638–4647. doi:10.1021/jm020503a

39. Chuaqui C, Deng Z, Singh J. Interaction profiles of protein kinase-inhibitor complexes and their application to virtual screening. J Med Chem. 2005;48: 121–133. doi:10.1021/jm049312t

40. Herzberg O, Moult J. Analysis of the steric strain in the polypeptide backbone of protein molecules. Proteins. 1991;11: 223–229. doi:10.1002/prot.340110307

41. Heringa J, Argos P. Strain in protein structures as viewed through nonrotameric side chains: II. effects upon ligand binding. Proteins. 1999;37: 44–55.

42. Laskowski RA, Luscombe NM, Swindells MB, Thornton JM. Protein clefts in molecular recognition and function. Protein Sci. 1996;5: 2438–2452. doi:10.1002/pro.5560051206

43. Zvelebil MJ, Barton GJ, Taylor WR, Sternberg MJ. Prediction of protein secondary structure and active sites using the alignment of homologous sequences. J Mol Biol. 1987;195: 957–961. doi:10.1016/0022-2836(87)90501-8

44. McPhalen CA, Vincent MG, Picot D, Jansonius JN, Lesk AM, Chothia C. Domain closure in mitochondrial aspartate aminotransferase. J Mol Biol. 1992;227: 197–213. doi:10.1016/0022-2836(92)90691-c

45. Irving JA, Whisstock JC, Lesk AM. Protein structural alignments and functional genomics. Proteins: Structure, Function, and Bioinformatics. 2001;42: 378–382. doi:https://doi.org/10.1002/1097-0134(20010215)42:3<378::AID-PROT70>3.0.CO;2-3

46. Leis S, Schneider S, Zacharias M. In Silico Prediction of Binding Sites on Proteins. CMC. 2010;17: 1550–1562. doi:10.2174/092986710790979944

47. Chiti F, Stefani M, Taddei N, Ramponi G, Dobson CM. Rationalization of the effects of mutations on peptide and protein aggregation rates. Nature. 2003;424: 805–808. doi:10.1038/nature01891

48. Pechmann S, Levy ED, Tartaglia GG, Vendruscolo M. Physicochemical principles that regulate the competition between functional and dysfunctional association of proteins. PNAS. 2009;106: 10159–10164. doi:10.1073/pnas.0812414106

49. Fernandez-Escamilla A-M, Rousseau F, Schymkowitz J, Serrano L. Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. Nat Biotechnol. 2004;22: 1302–1306. doi:10.1038/nbt1012

50. Chung J-L, Wang W, Bourne PE. High-throughput identification of interacting protein-protein binding sites. BMC Bioinformatics. 2007;8: 223. doi:10.1186/1471-2105-8-223

51. Liang J, Edelsbrunner H, Woodward C. Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. Protein Sci. 1998;7: 1884–1897.

52. Mattos C, Ringe D. Locating and characterizing binding sites on proteins. Nature Biotechnology. 1996;14: 595–599. doi:10.1038/nbt0596-595

53. Miller DW, Dill KA. Ligand binding to proteins: the binding landscape model. Protein Sci. 1997;6: 2166–2179.

54. Campbell SJ, Gold ND, Jackson RM, Westhead DR. Ligand binding: functional site location, similarity and docking. Curr Opin Struct Biol. 2003;13: 389–395. doi:10.1016/s0959-440x(03)00075-7

55. Vajda S, Guarnieri F. Characterization of protein-ligand interaction sites using experimental and computational methods. Curr Opin Drug Discov Devel. 2006;9: 354–362.

56. Burgoyne NJ, Jackson RM. Predicting protein interaction sites: binding hot-spots in protein–protein and protein–ligand interfaces. Bioinformatics. 2006;22: 1335–1342. doi:10.1093/bioinformatics/btl079

57. Laurie ATR, Jackson RM. Methods for the prediction of protein-ligand binding sites for structure-based drug design and virtual ligand screening. Curr Protein Pept Sci. 2006;7: 395–406. doi:10.2174/138920306778559386

58. Le Guilloux V, Schmidtke P, Tuffery P. Fpocket: An open source platform for ligand pocket detection. BMC Bioinformatics. 2009;10: 168. doi:10.1186/1471-2105-10-168

59. Schmidtke P, Barril X. Understanding and predicting druggability. A high-throughput method for detection of drug binding sites. J Med Chem. 2010;53: 5858–5867. doi:10.1021/jm100574m

60. Kozakov D, Hall DR, Napoleon RL, Yueh C, Whitty A, Vajda S. New Frontiers in Druggability. J Med Chem. 2015;58: 9063–9088. doi:10.1021/acs.jmedchem.5b00586

61. Agoni C, Olotu FA, Ramharack P, Soliman ME. Druggability and drug-likeness concepts in drug design: are biomodelling and predictive tools having their say? J Mol Model. 2020;26: 120. doi:10.1007/s00894-020-04385-6

62. Vistoli G, Pedretti A, Testa B. Assessing drug-likeness--what are we missing? Drug Discov Today. 2008;13: 285–294. doi:10.1016/j.drudis.2007.11.007

63. Hersey A, Chambers J, Bellis L, Patrícia Bento A, Gaulton A, Overington JP. Chemical databases: curation or integration by user-defined equivalence? Drug Discovery Today: Technologies. 2015;14: 17–24. doi:10.1016/j.ddtec.2015.01.005

64. Walters WP, Ajay null, Murcko MA. Recognizing molecules with drug-like properties. Curr Opin Chem Biol. 1999;3: 384–387. doi:10.1016/s1367-5931(99)80058-1

65. Walters WP, Stahl MT, Murcko MA. Virtual screening—an overview. Drug Discovery Today. 1998;3: 160–178. doi:10.1016/S1359-6446(97)01163-X

66. Oprea TI. Property distribution of drug-related chemical databases. J Comput Aided Mol Des. 2000;14: 251–264. doi:10.1023/a:1008130001697

67. Leeson PD, Springthorpe B. The influence of drug-like concepts on decision-making in medicinal chemistry. Nat Rev Drug Discov. 2007;6: 881–890. doi:10.1038/nrd2445

68. Kitchen DB, Decornez H, Furr JR, Bajorath J. Docking and scoring in virtual screening for drug discovery: methods and applications. Nature Reviews Drug Discovery. 2004;3: 935–949. doi:10.1038/nrd1549

69. Morris GM, Huey R, Lindstrom W, Sanner MF, Belew RK, Goodsell DS, et al. AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. J Comput Chem. 2009;30: 2785–2791. doi:10.1002/jcc.21256

70. Trott O, Olson AJ. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization and multithreading. J Comput Chem. 2010;31: 455–461. doi:10.1002/jcc.21334

71. H Z, A C. Discovery of ZAP70 inhibitors by high-throughput docking into a conformation of its kinase domain generated by molecular dynamics. Bioorg Med Chem Lett. 2013;23: 5721–5726. doi:10.1016/j.bmcl.2013.08.009

72. Allen WJ, Balius TE, Mukherjee S, Brozell SR, Moustakas DT, Lang PT, et al. DOCK 6: Impact of new features and current docking performance. J Comput Chem. 2015;36: 1132–1156. doi:10.1002/jcc.23905

73. Friesner RA, Banks JL, Murphy RB, Halgren TA, Klicic JJ, Mainz DT, et al. Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. Journal of Medicinal Chemistry. 2004;47: 1739–1749. doi:10.1021/jm0306430

74. Jones G, Willett P, Glen RC, Leach AR, Taylor R. Development and validation of a genetic algorithm for flexible docking. J Mol Biol. 1997;267: 727–748. doi:10.1006/jmbi.1996.0897

75. Corbeil CR, Williams CI, Labute P. Variability in docking success rates due to dataset preparation. J Comput Aided Mol Des. 2012;26: 775–786. doi:10.1007/s10822-012-9570-1

76. Spitzer R, Jain AN. Surflex-Dock: Docking Benchmarks and Real-World Application. J Comput Aided Mol Des. 2012;26: 687–699. doi:10.1007/s10822-011-9533-y

77. Chen Y-C. Beware of docking! Trends Pharmacol Sci. 2015;36: 78–95. doi:10.1016/j.tips.2014.12.001

78. Bello M, Martínez-Archundia M, Correa-Basurto J. Automated docking for novel drug discovery. Expert Opin Drug Discov. 2013;8: 821–834. doi:10.1517/17460441.2013.794780

79. Sousa SF, Ribeiro AJM, Coimbra JTS, Neves RPP, Martins SA, Moorthy NSHN, et al. Protein-ligand docking in the new millennium--a retrospective of 10 years in the field. Curr Med Chem. 2013;20: 2296–2314. doi:10.2174/0929867311320180002

80. Huang S-Y, Zou X. Advances and challenges in protein-ligand docking. Int J Mol Sci. 2010;11: 3016–3034. doi:10.3390/ijms11083016

81. Huang S-Y, Grinter SZ, Zou X. Scoring functions and their evaluation methods for protein-ligand docking: recent advances and future directions. Phys Chem Chem Phys. 2010;12: 12899–12908. doi:10.1039/c0cp00151a

82. Gohlke H, Hendlich M, Klebe G. Knowledge-based scoring function to predict protein-ligand interactions. J Mol Biol. 2000;295: 337–356. doi:10.1006/jmbi.1999.3371

83. Schulz-Gasch T, Stahl M. Scoring functions for protein-ligand interactions: a critical perspective. Drug Discov Today Technol. 2004;1: 231–239. doi:10.1016/j.ddtec.2004.08.004

84. Pecina A, Meier R, Fanfrlík J, Lepšík M, Řezáč J, Hobza P, et al. The SQM/COSMO filter: reliable native pose identification based on the quantum-mechanical description of protein–ligand interactions and implicit COSMO solvation. Chem Commun. 2016;52: 3312–3315. doi:10.1039/C5CC09499B

85. Raha K, Merz KM. Large-Scale Validation of a Quantum Mechanics Based Scoring Function: Predicting the Binding Affinity and the Binding Mode of a Diverse Set of Protein−Ligand Complexes. J Med Chem. 2005;48: 4558–4575. doi:10.1021/jm048973n

86. Moitessier N, Englebienne P, Lee D, Lawandi J, Corbeil CR. Towards the development of universal, fast and highly accurate docking/scoring methods: a long way to go. Br J Pharmacol. 2008;153 Suppl 1: S7-26. doi:10.1038/sj.bjp.0707515

87. Wang Z, Sun H, Yao X, Li D, Xu L, Li Y, et al. Comprehensive evaluation of ten docking programs on a diverse set of protein-ligand complexes: the prediction accuracy of sampling power and scoring power. Phys Chem Chem Phys. 2016;18: 12964–12975. doi:10.1039/c6cp01555g

88. Nissink JWM, Murray C, Hartshorn M, Verdonk ML, Cole JC, Taylor R. A new test set for validating predictions of protein-ligand interaction. Proteins. 2002;49: 457–471. doi:10.1002/prot.10232

89. Shoichet BK, McGovern SL, Wei B, Irwin JJ. Lead discovery using molecular docking. Curr Opin Chem Biol. 2002;6: 439–446. doi:10.1016/s1367-5931(02)00339-3

90. Aguirre-Alvarado C, Segura-Cabrera A, Velázquez-Quesada I, Hernández-Esquivel MA, García-Pérez CA, Guerrero-Rodríguez SL, et al. Virtual screening-driven repositioning of etoposide as CD44 antagonist in breast cancer cells. Oncotarget. 2016;7: 23772–23784. doi:10.18632/oncotarget.8180

91. Velázquez-Quesada I, Ruiz-Moreno AJ, Casique-Aguirre D, Aguirre-Alvarado C, Cortés-Mendoza F, de la Fuente-Granada M, et al. Pranlukast Antagonizes CD49f and Reduces Stemness in Triple-Negative Breast Cancer Cells. Drug Des Devel Ther. 2020;14: 1799–1811. doi:10.2147/DDDT.S247730

92. Knegtel RM, Kuntz ID, Oshiro CM. Molecular docking to ensembles of protein structures. J Mol Biol. 1997;266: 424–440. doi:10.1006/jmbi.1996.0776

93. Gohlke H, Klebe G. DrugScore Meets CoMFA: Adaptation of Fields for Molecular Comparison (AFMoC) or How to Tailor Knowledge-Based Pair-Potentials to a Particular Protein. J Med Chem. 2002;45: 4153–4170. doi:10.1021/jm020808p

94. Wu G, Vieth M. SDOCKER: A Method Utilizing Existing X-ray Structures To Improve Docking Accuracy. J Med Chem. 2004;47: 3142–3148. doi:10.1021/jm040015y

95. Ferrari AM, Wei BQ, Costantino L, Shoichet BK. Soft Docking and Multiple Receptor Conformations in Virtual Screening. J Med Chem. 2004;47: 5076–5084. doi:10.1021/jm049756p

96. Kovacs JA, Chacón P, Abagyan R. Predictions of protein flexibility: First-order measures. Proteins: Structure, Function, and Bioinformatics. 2004;56: 661–668. doi:https://doi.org/10.1002/prot.20151

97. Claussen H, Buning C, Rarey M, Lengauer T. FlexE: efficient molecular docking considering protein structure variations. J Mol Biol. 2001;308: 377–395. doi:10.1006/jmbi.2001.4551

98. Rarey M, Kramer B, Lengauer T. The particle concept: placing discrete water molecules during protein-ligand docking predictions. Proteins. 1999;34: 17–28.

99. Cohen NC, Tschinke V. Generation of new-lead structures in computer-aided drug design. Prog Drug Res. 1995;45: 205–243. doi:10.1007/978-3-0348-7164-8_6

100. Acharya C, Coop A, Polli JE, MacKerell AD. Recent Advances in Ligand-Based Drug Design: Relevance and Utility of the Conformationally Sampled Pharmacophore Approach. Curr Comput Aided Drug Des. 2011;7: 10–22.

101. Clark DE, Murray CW, Li J. Current Issues in De Novo Molecular Design. Reviews in Computational Chemistry. John Wiley & Sons, Ltd; 1997. pp. 67–125. doi:10.1002/9780470125885.ch2

102. Hajduk PJ, Greer J. A decade of fragment-based drug design: strategic advances and lessons learned. Nat Rev Drug Discov. 2007;6: 211–219. doi:10.1038/nrd2220

103. Rarey M. Molecular Design. Concepts and Applications. By Gisbert Schneider and Karl-Heinz Baringhaus. Angewandte Chemie International Edition. 2009;48: 1718–1719. doi:10.1002/anie.200900047

104. Gillet VJ, Khatib W, Willett P, Fleming PJ, Green DVS. Combinatorial Library Design Using a Multiobjective Genetic Algorithm. J Chem Inf Comput Sci. 2002;42: 375–385. doi:10.1021/ci010375j

105. Gillet VJ. New directions in library design and analysis. Current Opinion in Chemical Biology. 2008;12: 372–378. doi:10.1016/j.cbpa.2008.02.015

106. Hartenfeller M, Schneider G. De Novo Drug Design. In: Bajorath J, editor. Chemoinformatics and Computational Chemical Biology. Totowa, NJ: Humana Press; 2010. pp. 299–323. doi:10.1007/978-1-60761-839-3_12

107. Schneider G, Fechner U. Computer-based de novo design of drug-like molecules. Nat Rev Drug Discov. 2005;4: 649–663. doi:10.1038/nrd1799

108. Hopkins AL, Groom CR, Alex A. Ligand efficiency: a useful metric for lead selection. Drug Discov Today. 2004;9: 430–431. doi:10.1016/S1359-6446(04)03069-7

109. Bembenek SD, Tounge BA, Reynolds CH. Ligand efficiency and fragment-based drug discovery. Drug Discov Today. 2009;14: 278–283. doi:10.1016/j.drudis.2008.11.007

110. Schneider G, Lee ML, Stahl M, Schneider P. De novo design of molecular architectures by evolutionary assembly of drug-derived building blocks. J Comput Aided Mol Des. 2000;14: 487–494. doi:10.1023/a:1008184403558

111. Dolle RE. Comprehensive Survey of Combinatorial Library Synthesis: 2003. J Comb Chem. 2004;6: 623–679. doi:10.1021/cc0499082

112. Weber L. Current Status of Virtual Combinatorial Library Design. QSAR Comb Sci. 2005;24: 809–823. doi:10.1002/qsar.200510120

113. Weber L. The application of multi-component reactions in drug discovery. Curr Med Chem. 2002;9: 2085–2093. doi:10.2174/0929867023368719

114. Ruiz-Moreno AJ, Reyes-Romero A, Dömling A, Velasco-Velázquez MA. In Silico Design and Selection of New Tetrahydroisoquinoline-Based CD44 Antagonist Candidates. Molecules. 2021;26: 1877. doi:10.3390/molecules26071877

115. Saurabh S, Sivakumar PM, Perumal V, Khosravi A, Sugumaran A, Prabhawathi V. Molecular Dynamics Simulations in Drug Discovery and Drug Delivery. In: Krishnan A, Chuturgoon A, editors. Integrative Nanomedicine for New Therapies. Cham: Springer International Publishing; 2020. pp. 275–301. doi:10.1007/978-3-030-36260-7_10

116. Salo-Ahen OMH, Alanko I, Bhadane R, Bonvin AMJJ, Honorato RV, Hossain S, et al. Molecular Dynamics Simulations in Drug Discovery and Pharmaceutical Development. Processes. 2020;9: 71. doi:10.3390/pr9010071

117. Leach A. Molecular Modelling: Principles and Applications. Harlow, England ; New York; 2001.

118. Gunsteren WF van, Berendsen HJC. Computer Simulation of Molecular Dynamics: Methodology, Applications, and Perspectives in Chemistry. Angewandte Chemie International Edition in English. 1990;29: 992–1023. doi:https://doi.org/10.1002/anie.199009921

119. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic Acids Res. 2000;28: 235–242. doi:10.1093/nar/28.1.235

120. Zhou R. Replica exchange molecular dynamics method for protein folding simulation. Methods Mol Biol. 2007;350: 205–223. doi:10.1385/1-59745-189-4:205

121. Hamelberg D, Mongan J, McCammon JA. Accelerated molecular dynamics: A promising and efficient simulation method for biomolecules. J Chem Phys. 2004;120: 11919–11929. doi:10.1063/1.1755656

122. Laio A, Gervasio FL. Metadynamics: a method to simulate rare events and reconstruct the free energy in biophysics, chemistry and material science. Rep Prog Phys. 2008;71: 126601. doi:10.1088/0034-4885/71/12/126601

123. Pantsar T. The current understanding of KRAS protein structure and dynamics. Computational and Structural Biotechnology Journal. 2020;18: 189–198. doi:10.1016/j.csbj.2019.12.004

124. Prakash P, Gorfe AA. Lessons from computer simulations of Ras proteins in solution and in membrane. Biochimica et Biophysica Acta (BBA) - General Subjects. 2013;1830: 5211–5218. doi:10.1016/j.bbagen.2013.07.024

125. Pálfy G, Menyhárd DK, Perczel A. Dynamically encoded reactivity of Ras enzymes: opening new frontiers for drug discovery. Cancer Metastasis Rev. 2020;39: 1075–1089. doi:10.1007/s10555-020-09917-3

126. Dunker AK, Silman I, Uversky VN, Sussman JL. Function and structure of inherently disordered proteins. Curr Opin Struct Biol. 2008;18: 756–764. doi:10.1016/j.sbi.2008.10.002

127. Uversky VN. Intrinsically disordered proteins and novel strategies for drug discovery. Expert Opin Drug Discov. 2012;7: 475–488. doi:10.1517/17460441.2012.686489

128. Best RB. Computational and theoretical advances in studies of intrinsically disordered proteins. Curr Opin Struct Biol. 2017;42: 147–154. doi:10.1016/j.sbi.2017.01.006

129. Sinha N, Li Y, Lipschultz CA, Smith-Gill SJ. Understanding antibody–antigen associations by molecular dynamics simulations: Detection of important intra- and inter-molecular salt bridges. Cell Biochem Biophys. 2007;47: 361–375. doi:10.1007/s12013-007-0031-8

130. Yamashita T. Toward rational antibody design: recent advancements in molecular dynamics simulations. International Immunology. 2018;30: 133–140. doi:10.1093/intimm/dxx077

131. Kuroda D, Shirai H, Jacobson MP, Nakamura H. Computer-aided antibody design. Protein Eng Des Sel. 2012;25: 507–521. doi:10.1093/protein/gzs024

132. Li Z, Huang Y, Wu Y, Chen J, Wu D, Zhan C-G, et al. Absolute Binding Free Energy Calculation and Design of a Subnanomolar Inhibitor of Phosphodiesterase-10. J Med Chem. 2019;62: 2099–2111. doi:10.1021/acs.jmedchem.8b01763

133. Abel R, Wang L, Harder ED, Berne BJ, Friesner RA. Advancing Drug Discovery through Enhanced Free Energy Calculations. Acc Chem Res. 2017;50: 1625–1632. doi:10.1021/acs.accounts.7b00083

134. Huang YM, Chen W, Potter MJ, Chang CA. Insights from Free-Energy Calculations: Protein Conformational Equilibrium, Driving Forces, and Ligand-Binding Modes. Biophysical Journal. 2012;103: 342–351. doi:10.1016/j.bpj.2012.05.046

135. Ezerski JC, Zhang P, Jennings NC, Waxham MN, Cheung MS. Molecular Dynamics Ensemble Refinement of Intrinsically Disordered Peptides According to Deconvoluted Spectra from Circular Dichroism. Biophysical Journal. 2020;118: 1665–1678. doi:10.1016/j.bpj.2020.02.015

136. Cuendet MA, Zoete V, Michielin O. How T cell receptors interact with peptide-MHCs: a multiple steered molecular dynamics study. Proteins. 2011;79: 3007–3024. doi:10.1002/prot.23104

137. Morrone JA, Perez A, Deng Q, Ha SN, Holloway MK, Sawyer TK, et al. Molecular Simulations Identify Binding Poses and Approximate Affinities of Stapled α-Helical Peptides to MDM2 and MDMX. J Chem Theory Comput. 2017;13: 863–869. doi:10.1021/acs.jctc.6b00978

138. Lelimousin M, Limongelli V, Sansom MSP. Conformational Changes in the Epidermal Growth Factor Receptor: Role of the Transmembrane Domain Investigated by Coarse-Grained MetaDynamics Free Energy Calculations. J Am Chem Soc. 2016;138: 10611–10622. doi:10.1021/jacs.6b05602

139. Mustafa G, Nandekar PP, Yu X, Wade RC. On the application of the MARTINI coarse-grained model to immersion of a protein in a phospholipid bilayer. J Chem Phys. 2015;143: 243139. doi:10.1063/1.4936909

140. Ash WL, Zlomislic MR, Oloo EO, Tieleman DP. Computer simulations of membrane proteins. Biochimica et Biophysica Acta (BBA) - Biomembranes. 2004;1666: 158–189. doi:10.1016/j.bbamem.2004.04.012

141. Kandt C, Ash WL, Tieleman DP. Setting up and running molecular dynamics simulations of membrane proteins. Methods. 2007;41: 475–488. doi:10.1016/j.ymeth.2006.08.006

142. Larsen AS, Ruggiero MT, Johansson KE, Zeitler JA, Rantanen J. Tracking Dehydration Mechanisms in Crystalline Hydrates with Molecular Dynamics Simulations. Crystal Growth & Design. 2017;17: 5017–5022. doi:10.1021/acs.cgd.7b00889

143. Knapik J, Wojnarowska Z, Grzybowska K, Tajber L, Mesallati H, Paluch KJ, et al. Molecular Dynamics and Physical Stability of Amorphous Nimesulide Drug and Its Binary Drug-Polymer Systems. Mol Pharm. 2016;13: 1937–1946. doi:10.1021/acs.molpharmaceut.6b00115

144. Khalkhali M, Mohammadinejad S, Khoeini F, Rostamizadeh K. Vesicle-like structure of lipid-based nanoparticles as drug delivery system revealed by molecular dynamics simulations. International Journal of Pharmaceutics. 2019;559: 173–181. doi:10.1016/j.ijpharm.2019.01.036

145. Mak K-K, Pichika MR. Artificial intelligence in drug development: present status and future prospects. Drug Discov Today. 2019;24: 773–780. doi:10.1016/j.drudis.2018.11.014

146. Singh B. Artificial Intelligence: The Beginning of a New Era in Pharmacy Profession. Asian Journal of Pharmaceutics. 2018;12: 72–76.

147. Paul D, Sanap G, Shenoy S, Kalyane D, Kalia K, Tekade RK. Artificial intelligence in drug discovery and development. Drug Discov Today. 2021;26: 80–93. doi:10.1016/j.drudis.2020.10.010

148. Smith RG, Farquhar A. The Road Ahead for Knowledge Management: An AI Perspective. AIMag. 2000;21: 17–17. doi:10.1609/aimag.v21i4.1528

149. Lamberti MJ, Wilkinson M, Donzanti BA, Wohlhieter GE, Parikh S, Wilkins RG, et al. A Study on the Application and Use of Artificial Intelligence to Support Drug Development. Clin Ther. 2019;41: 1414–1426. doi:10.1016/j.clinthera.2019.05.018

150. Beneke F, Mackenrodt M-O. Artificial Intelligence and Collusion. IIC. 2019;50: 109–134. doi:10.1007/s40319-018-00773-x

151. Walczak S, Cerpa N. Artificial Neural Networks. In: Meyers RA, editor. Encyclopedia of Physical Science and Technology (Third Edition). New York: Academic Press; 2003. pp. 631–645. doi:10.1016/B0-12-227410-5/00837-1

152. Bielecki A. Models of Neurons and Perceptrons: Selected Problems and Challenges. Cham: Springer International Publishing; 2019. doi:10.1007/978-3-319-90140-4

153. Zang Q, Mansouri K, Williams AJ, Judson RS, Allen DG, Casey WM, et al. In Silico Prediction of Physicochemical Properties of Environmental Chemicals Using Molecular Fingerprints and Machine Learning. J Chem Inf Model. 2017;57: 36–49. doi:10.1021/acs.jcim.6b00625

154. Yang X, Wang Y, Byrne R, Schneider G, Yang S. Concepts of Artificial Intelligence for Computer-Assisted Drug Discovery. Chem Rev. 2019;119: 10520–10594. doi:10.1021/acs.chemrev.8b00728

155. Hessler G, Baringhaus K-H. Artificial Intelligence in Drug Design. Molecules. 2018;23: 2520. doi:10.3390/molecules23102520

156. Öztürk H, Özgür A, Ozkirimli E. DeepDTA: deep drug-target binding affinity prediction. Bioinformatics. 2018;34: i821–i829. doi:10.1093/bioinformatics/bty593

157. Nascimento ACA, Prudêncio RBC, Costa IG. A multiple kernel learning algorithm for drug-target interaction prediction. BMC Bioinformatics. 2016;17: 46. doi:10.1186/s12859-016-0890-3

158. Tyrchan C, Evertsson E. Matched Molecular Pair Analysis in Short: Algorithms, Applications and Limitations. Computational and Structural Biotechnology Journal. 2017;15: 86–90. doi:10.1016/j.csbj.2016.12.003

159. Degen J, Wegscheid-Gerlach C, Zaliani A, Rarey M. On the Art of Compiling and Using "Drug-Like" Chemical Fragment Spaces. ChemMedChem. 2008;3: 1503–1507. doi:10.1002/cmdc.200800178

160. Pereira JC, Caffarena ER, dos Santos CN. Boosting Docking-Based Virtual Screening with Deep Learning. J Chem Inf Model. 2016;56: 2495–2506. doi:10.1021/acs.jcim.6b00355

161. Sheridan RP, Wang WM, Liaw A, Ma J, Gifford EM. Extreme Gradient Boosting as a Method for Quantitative Structure–Activity Relationships. J Chem Inf Model. 2016;56: 2353–2360. doi:10.1021/acs.jcim.6b00591

162. Wallach I, Dzamba M, Heifets A. AtomNet: A Deep Convolutional Neural Network for Bioactivity Prediction in Structure-based Drug Discovery. arXiv:151002855 [cs, q-bio, stat]. 2015 [cited 11 Jun 2021]. Available: http://arxiv.org/abs/1510.02855

163. Turk S, Merget B, Rippmann F, Fulle S. Coupling Matched Molecular Pairs with Machine Learning for Virtual Compound Optimization. J Chem Inf Model. 2017;57: 3079–3085. doi:10.1021/acs.jcim.7b00298

164. Leach AG, Jones HD, Cosgrove DA, Kenny PW, Ruston L, MacFaul P, et al. Matched Molecular Pairs as a Guide in the Optimization of Pharmaceutical Properties; a Study of Aqueous Solubility, Plasma Protein Binding and Oral Exposure. J Med Chem. 2006;49: 6672–6682. doi:10.1021/jm0605233

165. Keefer CE, Chang G, Kauffman GW. Extraction of tacit knowledge from large ADME data sets via pairwise analysis. Bioorganic & Medicinal Chemistry. 2011;19: 3739–3749. doi:10.1016/j.bmc.2011.05.003

166. Warner DJ, Griffen EJ, St-Gallay SA. WizePairZ: A Novel Algorithm to Identify, Encode, and Exploit Matched Molecular Pairs with Unspecified Cores in Medicinal Chemistry. J Chem Inf Model. 2010;50: 1350–1357. doi:10.1021/ci100084s

167. Dossetter AG. A statistical analysis of in vitro human microsomal metabolic stability of small phenyl group substituents, leading to improved design sets for parallel SAR exploration of a chemical series. Bioorganic & Medicinal Chemistry. 2010;18: 4405–4414. doi:10.1016/j.bmc.2010.04.077

168. Schyman P, Liu R, Desai V, Wallqvist A. vNN Web Server for ADMET Predictions. Front Pharmacol. 2017;8. doi:10.3389/fphar.2017.00889

169. Schönherr H, Cernak T. Profound Methyl Effects in Drug Discovery and a Call for New C□H Methylation Reactions. Angewandte Chemie International Edition. 2013;52: 12256–12267. doi:10.1002/anie.201303207

170. Mayr A, Klambauer G, Unterthiner T, Hochreiter S. DeepTox: Toxicity Prediction using Deep Learning. Front Environ Sci. 2016;3. doi:10.3389/fenvs.2015.00080

171. Wan F, Zeng J (Michael). Deep learning with feature embedding for compound-protein interaction prediction. bioRxiv. 2016; 086033. doi:10.1101/086033

172. AlphaFold: a solution to a 50-year-old grand challenge in biology. In: Deepmind [Internet]. [cited 15 Apr 2021]. Available: /blog/article/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biology

173. Tian K, Shao M, Wang Y, Guan J, Zhou S. Boosting compound-protein interaction prediction by deep learning. Methods. 2016;110: 64–72. doi:10.1016/j.ymeth.2016.06.024

174. Wang F, Liu D, Wang H, Luo C, Zheng M, Liu H, et al. Computational screening for active compounds targeting protein sequences: methodology and experimental validation. J Chem Inf Model. 2011;51: 2821–2828. doi:10.1021/ci200264h

175. Yu H, Chen J, Xu X, Li Y, Zhao H, Fang Y, et al. A systematic prediction of multiple drug-target interactions from chemical, genomic, and pharmacological data. PLoS One. 2012;7: e37608. doi:10.1371/journal.pone.0037608

176. Chadderton M. Drug repositioning_Layout 1. Drug Development. 2011; 4.

177. Li X, Xu Y, Cui H, Huang T, Wang D, Lian B, et al. Prediction of synergistic anti-cancer drug combinations based on drug target network and drug induced gene expression profiles. Artif Intell Med. 2017;83: 35–43. doi:10.1016/j.artmed.2017.05.008

178. Reddy AS, Zhang S. Polypharmacology: drug discovery for the future. Expert Rev Clin Pharmacol. 2013;6: 41–47. doi:10.1586/ecp.12.74

179. Achenbach J, Tiikkainen P, Franke L, Proschak E. Computational tools for polypharmacology and repurposing. Future Med Chem. 2011;3: 961–968. doi:10.4155/fmc.11.62

180. Li Z, Li X, Liu X, Fu Z, Xiong Z, Wu X, et al. KinomeX: a web application for predicting kinome-wide polypharmacology effect of small molecules. Bioinformatics. 2019;35: 5354–5356. doi:10.1093/bioinformatics/btz519

181. Cyclica Launches Ligand Express™, a Disruptive Cloud-Based Platform to Revolutionize Drug Discovery — Cyclica. [cited 15 Apr 2021]. Available: https://www.cyclicarx.com/press-releases/cyclica-launches-ligand-express-a-disruptive-cloud-based-platform-to-revolutionize-drug-discovery

182. Hayik SA, Dunbrack R, Merz KM. Mixed Quantum Mechanics/Molecular Mechanics Scoring Function To Predict Protein−Ligand Binding Affinity. J Chem Theory Comput. 2010;6: 3079–3091. doi:10.1021/ct100315g

183. Ryde U. Chapter Six - QM/MM Calculations on Proteins. In: Voth GA, editor. Methods in Enzymology. Academic Press; 2016. pp. 119–158. doi:10.1016/bs.mie.2016.05.014

184. Smith JS, Isayev O, Roitberg AE. ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost. Chem Sci. 2017;8: 3192–3203. doi:10.1039/C6SC05720A

185. Zhang Y-J, Khorshidi A, Kastlunger G, Peterson AA. The potential for machine learning in hybrid QM/MM calculations. J Chem Phys. 2018;148: 241740. doi:10.1063/1.5029879

186. Faber FA, Lindmaa A, von Lilienfeld OA, Armiento R. Machine Learning Energies of 2 Million Elpasolite $(AB{C}_{2}{D}_{6})$ Crystals. Phys Rev Lett. 2016;117: 135502. doi:10.1103/PhysRevLett.117.135502

187. Corey EJ, Wipke WT. Computer-assisted design of complex organic syntheses. Science. 1969;166: 178–192. doi:10.1126/science.166.3902.178

188. Collins KD, Glorius F. A robustness screen for the rapid assessment of chemical reactions. Nature Chem. 2013;5: 597–601. doi:10.1038/nchem.1669

189. Segler MHS, Preuss M, Waller MP. Planning chemical syntheses with deep neural networks and symbolic AI. Nature. 2018;555: 604–610. doi:10.1038/nature25978

190. Browne CB, Powley E, Whitehouse D, Lucas SM, Cowling PI, Rohlfshagen P, et al. A Survey of Monte Carlo Tree Search Methods. IEEE Transactions on Computational Intelligence and AI in Games. 2012;4: 1–43. doi:10.1109/TCIAIG.2012.2186810

191. Kayala MA, Azencott C-A, Chen JH, Baldi P. Learning to Predict Chemical Reactions. J Chem Inf Model. 2011;51: 2209–2222. doi:10.1021/ci200207y

192. Cook A, Johnson AP, Law J, Mirzazadeh M, Ravitz O, Simon A. Computer-aided synthesis design: 40 years on. WIREs Computational Molecular Science. 2012;2: 79–107. doi:10.1002/wcms.61

193. Segler MHS, Waller MP. Neural-Symbolic Machine Learning for Retrosynthesis and Reaction Prediction. Chemistry – A European Journal. 2017;23: 5966–5971. doi:10.1002/chem.201605499

194. Chan HCS, Shan H, Dahoun T, Vogel H, Yuan S. Advancing Drug Discovery via Artificial Intelligence. Trends Pharmacol Sci. 2019;40: 592–604. doi:10.1016/j.tips.2019.06.004

195. Grzybowski BA, Szymkuć S, Gajewska EP, Molga K, Dittwald P, Wołos A, et al. Chematica: A Story of Computer Code That Started to Think like a Chemist. Chem. 2018;4: 390–398. doi:10.1016/j.chempr.2018.02.024

196. Klucznik T, Mikulak-Klucznik B, McCormack MP, Lima H, Szymkuć S, Bhowmick M, et al. Efficient Syntheses of Diverse, Medicinally Relevant Targets Planned by Computer and Executed in the Laboratory. Chem. 2018;4: 522–532. doi:10.1016/j.chempr.2018.02.002

197. Segler MHS, Kogej T, Tyrchan C, Waller MP. Generating Focused Molecule Libraries for Drug Discovery with Recurrent Neural Networks. ACS Cent Sci. 2018;4: 120–131. doi:10.1021/acscentsci.7b00512

198. Schneider G, Clark DE. Automated De Novo Drug Design: Are We Nearly There Yet? Angewandte Chemie International Edition. 2019;58: 10792–10803. doi:https://doi.org/10.1002/anie.201814681

199. Ripphausen P, Nisius B, Peltason L, Bajorath J. Quo vadis, virtual screening? A comprehensive survey of prospective applications. J Med Chem. 2010;53: 8461–8467. doi:10.1021/jm101020z

200. Damm-Ganamet KL, Bembenek SD, Venable JW, Castro GG, Mangelschots L, Peeters DCG, et al. A Prospective Virtual Screening Study: Enriching Hit Rates and Designing Focus Libraries To Find Inhibitors of PI3Kδ and PI3Kγ. J Med Chem. 2016;59: 4302–4313. doi:10.1021/acs.jmedchem.5b01974

201. Amaro RE, Li WW. Emerging methods for ensemble-based virtual screening. Curr Top Med Chem. 2010;10: 3–13. doi:10.2174/156802610790232279

202. Singh N, Chevé G, Ferguson DM, McCurdy CR. A combined ligand-based and target-based drug design approach for G-protein coupled receptors: application to salvinorin A, a selective kappa opioid receptor agonist. J Comput Aided Mol Des. 2006;20: 471–493. doi:10.1007/s10822-006-9067-x

203. Prathipati P, Mizuguchi K. Integration of Ligand and Structure Based Approaches for CSAR-2014. J Chem Inf Model. 2016;56: 974–987. doi:10.1021/acs.jcim.5b00477

204. Bauer JF. Polymorphism—A Critical Consideration in Pharmaceutical Development, Manufacturing, and Stability. : 9.

205. Fry DC. Protein-protein interactions as targets for small molecule drug discovery. Biopolymers. 2006;84: 535–552. doi:10.1002/bip.20608

206. Rantanen J, Khinast J. The Future of Pharmaceutical Manufacturing Sciences. J Pharm Sci. 2015;104: 3612–3638. doi:10.1002/jps.24594

207. Jämsä-Jounela ProfS-L. FUTURE TRENDS IN PROCESS AUTOMATION. IFAC Proceedings Volumes. 2007;40: 1–10. doi:10.3182/20070213-3-CU-2913.00003

Chapter 2

In Silico Design and Selection of New Tetrahydroisoquinoline-Based CD44 Antagonist Candidates

Angel J. Ruiz-Moreno, Atilio Reyes-Romero, Alexander Dömling, and Marco A. Velasco-Velázquez

This chapter has been published in Molecules 2021, 26, 1877.

Abstract

CD44 promotes metastasis, chemoresistance, and stemness in different types of cancer and is a target for the development of new anti-cancer therapies. All CD44 isoforms share a common N-terminal domain that binds to hyaluronic acid (HA). Herein, we used a computational approach to design new potential CD44 antagonists and evaluate their target-binding ability. By analyzing 30 crystal structures of the HA-binding domain (CD44HAbd), we characterized a subdomain that binds to 1,2,3,4-tetrahydroisoquinoline (THQ)-containing compounds and is adjacent to residues essential for HA interaction. By computational combinatorial chemistry (CCC), we designed 168,190 molecules and compared their conformers to a pharmacophore containing the key features of the crystallographic THQ binding mode. Approximately 0.01% of the compounds matched the pharmacophore and were analyzed by computational docking and molecular dynamics (MD). We identified two compounds, Can125 and Can159, that bound to human CD44HAbd (hCD44HAbd) in explicit-solvent MD simulations and therefore may elicit CD44 blockage. These compounds can be easily synthesized by multicomponent reactions for activity testing and their binding mode, reported here, could be helpful in the design of more potent CD44 antagonists.

Introduction

CD44 is a transmembrane glycoprotein that functions as a receptor for the glycosaminoglycan hyaluronic acid (HA), an integral component of the extracellular matrix [1,2]. CD44 is expressed on multiple cells, including embryonic stem cells and differentiated cells, mediating cellular functions such as adhesion, homing, migration, and extravasation [1,2]. CD44 transcript can undergo alternative splicing, generating multiple isoforms of CD44, but all of them conserve intact the HA-binding domain (HAbd) and, therefore, can be activated by HA [3].

CD44 expression correlates with unfavorable clinical outcomes in multiple types of cancer [4–8]. CD44 activation by HA in cancer cells induces transcriptional and epigenetic changes that stimulate signaling pathways controlling invasiveness and metastasis, chemoresistance, and stemness [9–11]. For instance, in breast cancer cells, HA binding to CD44 induces epithelial–mesenchymal transition, which increases cell migration and invasive capacity [12], and promotes survival under detached conditions during the development of metastasis [13]. Moreover, CD44 is expressed in cancer stem cells that survive chemotherapy in models of glioblastoma [14], breast [15], pancreatic [16], colorectal [17], and prostate [18] cancer. Consistent with its key role in cancer progression, CD44 silencing impairs chemoresistance, clonogenicity, tumorigenicity, and/or metastasis [19–21]. Therefore, blockage of HA-binding to CD44 has been proposed as a potential therapeutic strategy for cancer.

The CD44HAbd is in the N-terminal end of the extracellular region of the receptor. Structural analysis of murine CD44HAbd crystals showed that only 13 residues along a shallow groove mediate HA-binding [22]. The residues Arg41, Tyr42, Arg78, Tyr79 in hCD44HAbd (Arg45, Tyr46, Arg82, Tyr83 in mCD44HAbd) have been previously described as essential for HA-binding by directed mutagenesis experiments or crystal analysis [23,24]. Given the lack of an obvious druggable pocket in the HA-binding site, small molecule inhibitors that interact with allosteric sites within the CD44HAbd have been developed [24–27]. However, those compounds bind to CD44HAbd in the high micromolar or even low millimolar range, limiting further applications. Therefore, there is a need for new CD44 antagonists with improved affinity, efficacy, and physicochemical properties for future effective translation to the clinic.

Herein we designed and evaluated the binding of new potential CD44 antagonists using an *in silico* strategy. We identified that small molecules sharing a 1,2,3,4-tetrahydroisoquinoline (THQ)

motif are frequently co-crystallized with CD44HAbd in a subdomain adjacent to the residues that are essential for HA-binding. By computational combinatorial chemistry (CCC), we generated libraries including more than 168,000 THQ-containing molecules. The new molecules (i) could be easily synthesized by multicomponent reactions, (ii) are diverse, and (iii) display drug-like physicochemical properties. We selected a subset of 163 candidates matching the key features of the reported THQ binding mode for further analysis by computational docking. The nine candidates with the highest frequency of poses reproducing the reported THQ binding mode were analyzed by molecular dynamics (MD). Our results allowed the identification of two compounds predicted to stably bind to hCD44HAbd in an aqueous solution. Those compounds may be useful as CD44 antagonists, and the information of their binding mode can be employed as the basis for the design of new bioactive molecules that target CD44.

Materials and Methods

Sequence and Structural Alignments

Human and murine CD44 binding-domain (hCD44HAbd and mCD44HAbd, respectively) crystal structures were retrieved from the Protein Data Bank (PDB). All structures were aligned using UCSF Chimera 1.14 [28] and the A chain of the entry 1UUH as reference. RMSD was calculated for alpha-carbons, backbone, and all atoms. Sequence alignments were performed using the pairwise2 module of Biopython 1.78 [29]. For small-molecule atom-based alignment and comparison, a python tailored-made script was created employing the Cheminformatics Toolbox RDKIT (http://www.rdkit.org, accessed date 25 February 2021) and the Pymol 2.4 API (PyMOL Molecular Graphics System, Schrödinger, LLC). The script is available at https://github.com/AngelRuizMoreno/CD44_antagonist (accessed date 25 February 2021).

The 3D Pharmacophore Modeling

The 3D pharmacophoric model was generated by using 21 mCD44HAbd crystal structures containing small molecules with a THQ motif (Table S1) and the Pharmit server [30]. The most relevant molecular features were chosen by visual inspection, and their 3D coordinates in the mCD44HAbd pocket were set using the threshold of RMSD < 0.5 Å among THQ atoms.

Combinatorial Computational Chemistry

The creation of compound libraries by the CCC approach was carried out using Reactor 20.17 from ChemAxon (http://www.chemaxon.com, accessed date 25 February 2021). The Ugi Tetrazole and Ugi 3 component reactions were selected as synthesis routes to generate compounds synthetically accessible by MCR. For the CCC experiments, libraries of building blocks were made by searching highly diverse and low-cost commercially available starting materials using the sci-finder platform (https://scifinder.cas.org). The building blocks library consisted of 32 substituted 1,2,3,4-THQ, 4 aldehydes, and 657 isocyanides.

Cheminformatic Analysis

The resulting compounds from CCC were stored in two different libraries according to their synthesis origin. For comparison, the DrugBank dataset was included in the cheminformatic analysis. For each library, we computed tSNE, NPR, PCA, and Ro5. The tSNE analysis was computed using the Tanimoto similarity among the MACCS keys for each molecule using RDKIT; then, the tSNE calculation was conducted using Scikit learn 0.23.1 [31], implementing two components and a perplexity value of 50. After tSNE analysis, a silhouette-based k-means clustering was performed using Scikit learn 0.23.1 [31]. Similarly, the PCA was performed using Scikit learn and employing ten selected non-redundant molecular descriptors of 30 different 2D and 3D molecular descriptors, which were computed using RDKIT. The NPR analysis was carried out by calculating the NPR1 and NPR2 descriptors for the molecules. Finally, the Ro5 analysis was performed by calculating the molecular weight, LogP, number of hydrogen bond donors and acceptors, and TPSA.

The 3D Pharmacophoric Matching

We generated 20 energetically favorable conformers for each compound within our working libraries by using the Mcnf module of Moloc [32]. All the conformers were aligned against the THQ-atoms and pharmacophore using the Pharmit server [30]. The best conformers were selected using an RMSD threshold of 0.5 Å against the pharmacophoric model descriptors. Finally, a local optimization was performed using the mCD44HAbd surface using Moloc [32,33], followed by visual inspection. The molecules that after local optimization kept the 3D pharmacophoric matching (RMSD < 0.5 Å) were selected for further studies.

Molecular Docking

According to the identity and RMSD values, we found a very high similarity between all hCD44bd. Therefore, we choose the first hCD44bd crystal reported (1UUH). Regarding the mCD44HAbd our data also indicated a very high similarity among all available structures. Nonetheless, for docking experiments, we focused on the mCD44HAbd containing THQ-molecules, and we choose the crystal 5BZM, which contains a THQ-molecule displaying all molecular features according to our pharmacophoric model.

For the validation of the molecular docking protocol, the crystal structure of mCD44HAbd (5BZM) was used as the receptor and 21 co-crystallized THQ-containing molecules as ligands. For virtual screening, proteins hCD44HAbd (1UUH) and mCD44HAbd (5BZM) were employed as receptors, and the 163 molecules selected by pharmacophore filtering as ligands. Secondary dockings were carried out into additional pockets identified Fpocket v3.0 [34] in ligand-free hCD44HAbd (1UUH), HA-bound mCD44HAbd (2JCR), or THQ-containing molecule mCD44HAbd (5BZM).

Proteins were prepared by removing co-crystallized waters, solvent molecules, and adding charges and hydrogens using Chimera 1.14 [28]. Ligands were prepared by adding explicit hydrogens and tautomeric states at pH 7.4. and generating 3D coordinates with Standardized 19.20.0 (http://www.chemaxon.com, accessed date 25 February 2021). For virtual screening, docking was performed into hCD44HAbd and mCD44HAbd within a sphere with 8 Å of radius and center in the Thr27 or Thr31, respectively. Secondary dockings were carried out using as reference the coordinates of each additional pocket identified. For each ligand, 50 runs of the genetic algorithm were performed for the conformational search. Each pose was evaluated employing the PLP Chemscore scoring function established in the GOLD software from the Cambridge Crystallographic Data Center (CCDC) [35]. For each compound, the best 25 poses were saved for analysis. Finally, hierarchical clustering analysis of the poses was performed using Scipy 1.5.2 [36].

Molecular Dynamics

The MD simulations were carried out using Gromacs 5.0.4 [37]. Selected candidates and hCD44HAbd (1UUH) were parameterized using the CGenFF and CHARMM36 force field, respectively, through the CHARMM-GUI (http://www.charmm-gui.org/, accessed date 25 February 2021) [38]. The systems were built by adding TIP3P water molecules to the ligand-hCD44HAbd complexes, neutralizing ions, and establishing periodic boundary conditions (PBC)

by using the multicomponent assembler of the CHARMM-GUI. Before production, the systems were minimized and then equilibrated under an NVT assembly. During the production phase, an NPT assembly was performed at 310.15 K for 100 ns saving velocities, energy, and positions every ten ps. Water molecules displacement was computed by quantifying the number of water molecules within a 6 Å radius sphere covering the THQ-binding pocket, every frame of the simulation. Analysis of ligand-target interactions was carried out by a python tailored-made script (https://github.com/AngelRuizMoreno/CD44_antagonist, accessed date 25 February 2021) implementing MDAnalysis [39] and PLIP [40].

Interactions Free Energy Calculations

Full-length candidate-hCD44HAbd trajectories were employed for free energy calculations using the molecular mechanics energies combined with the Poisson–Boltzmann surface area continuum solvation (MM/PBSA) [41] by g_mmpbsa v1.6 package [42]. Computation of the potential energy in vacuum, polar solvation energy, and non-polar solvation energy were performed to calculate the average binding energy. Per-residue energetic decomposing and maps were created to show the contribution of each residue to the binding energy.

Results

Identification of a Target Subdomain within the CD44HAbd and Generation of a THQ-Based Pharmacophore

Aiming to identify relevant regions for drug design, we compared the 30 crystal structures available in Protein Data Bank (PDB) that comprise the HA-binding domain of human (three structures) or mouse (27 structures) CD44 (Table S1). The three human structures correspond to the apo form of CD44HAbd. For mCD44HAbd, 2JCP represents the apo-CD44HAbd, three structures are co-crystallized with HA (2JCQ, 2JCR, and 4MRD), and the rest are co-crystallized with molecules weighting 100–250 Da. Within the structures containing small molecules, 21 of them are co-crystallized with compounds containing the THQ motif. Sequence identity analysis showed 100% identity among all hCD44HAbd, 99–100% among mCD44HAbd, and 86–88% between hCD44HAbd and mCD44HAbd (Figure S1A). Due to the high identity among CD44HAbd structures, we compared all of them in a structural analysis. The root mean square deviation (RMSD) profiles for alpha-carbons and full atoms showed the higher deviations on some residues previously reported as essential for HA-binding by direct mutagenesis experiments (Arg41, Tyr42, Arg78, and Tyr79) in hCD44HAbd [23] (Figure S1B).

Analysis of the 21 structures co-crystallized with THQ-containing small molecules identified that all these ligands bind to a pocket contiguous to the HA-binding domain (Figure 1A). Their binding drives a shift in multiple residues of mCD44HAbd, including some of the key residues participating in HA-binding, namely, Arg45, Arg82, and Arg155 (Figure 1B). Alignment of THQ-composing atoms showed that the binding mode is highly conserved among the 21 crystals analyzed (Figure 1C). Thus, we model a pharmacophore from the co-crystallized molecules containing the THQ substructure. The generated model included four pharmacophoric descriptors—aromatic, two hydrogen bond donors, and a positively charged ion (Figure 1D).

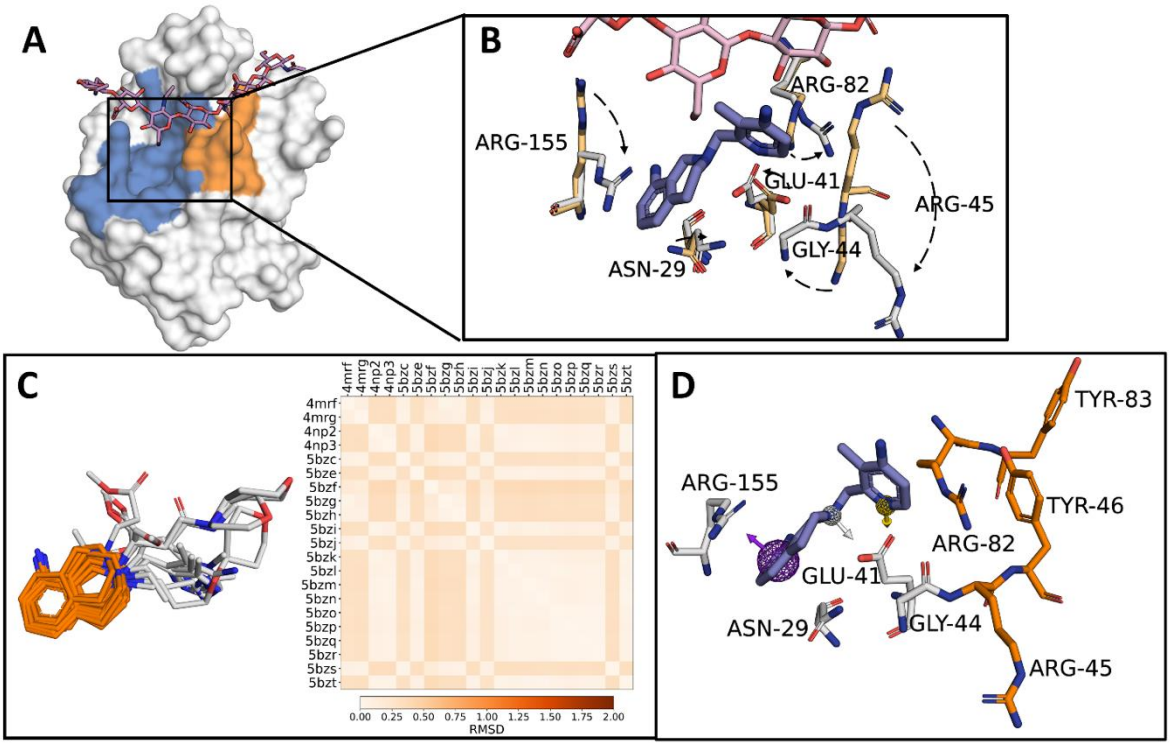
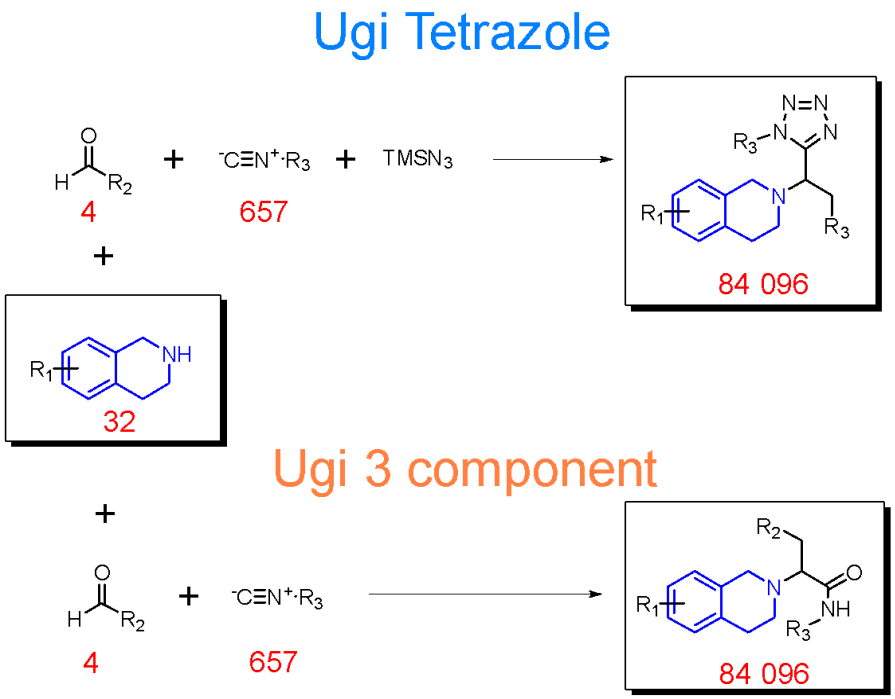


Figure 1. Generation of a 1,2,3,4-tetrahydroisoquinoline (THQ)-based pharmacophore. (**A**) CD44HAbd (white surface) has a subdomain where the THQ-containing molecules can bind (blue) and partially overlaps with the region containing the key residues for CD44–HA binding (Arg45, Tyr46, Arg82, and Tyr83-orange-). HA is shown in pink. (**B**) Spatial positions of the lateral chain of residues Asn29, Glu41, Gly44, Arg45, Arg82, and Arg155 from mCD44HAbd in the HA-bound form (residues shown as beige sticks; HA in pink) vs. the form bound to a THQ-containing molecule (residues shown as grey sticks; ligand in blue). Dashes represent the residue shifts between both states. (**C**) Structural alignment and root mean square deviation (RMSD) comparison of the THQ scaffold (orange) of the molecules co-crystallized with mCD44HAbd. (**D**) Pharmacophore model generated using THQ-containing molecules as a template. Purple sphere: aromatic; yellow sphere: hydrogen bond donor; white sphere: merged hydrogen bond donor and positively charged ion. A molecule with the THQ substructure is included (blue) to show the interacting residues in CD44 (grey sticks) and its proximity to residues essential for HA-binding (orange sticks).

Generation of THQ-Containing Libraries

To identify new compounds with the potential capability to interfere with the CD44–HA binding, we employed CCC to generate two libraries of compounds that include the THQ scaffold. To facilitate the synthesis of our compounds in subsequent research, we decided to use multicomponent reactions (MCR) synthesis routes. Considering the characteristics of the THQ-containing small molecules co-crystallized with mCD44HAbd, we implemented the Ugi four-component tetrazole synthesis [43,44] and Ugi three-component reaction [45] for our CCC experiments. For each MCR route, we obtained 84,096 different compounds (Scheme 1).



Scheme 1. CCC strategy for the generation of the test libraries.

Cheminformatic Analysis of CCC Compounds

In order to explore the chemical diversity and physicochemical properties of the designed compounds, we employed a series of cheminformatic analyses, calculating 30 different 2D/3D-shape and physicochemical molecular descriptors. For comparison, we also studied the compounds within DrugBank Database 5.0.10, a library of 1542 FDA-approved small molecules [46]. The diversity analysis of 1500 randomly sampled compounds of each library, using t-distributed stochastic neighbor embedding (tSNE) employing Molecular ACCess System (MACCS) keys [47], showed that the compounds from the tetrazole and Ugi libraries display similar structural diversity to the compounds inside DrugBank. The K-means clustering showed

that compounds from the CCC and DrugBank libraries distribute similarly on five out of six clusters, whereas the sixth cluster was enriched in DrugBank small molecules (Figure 2A).

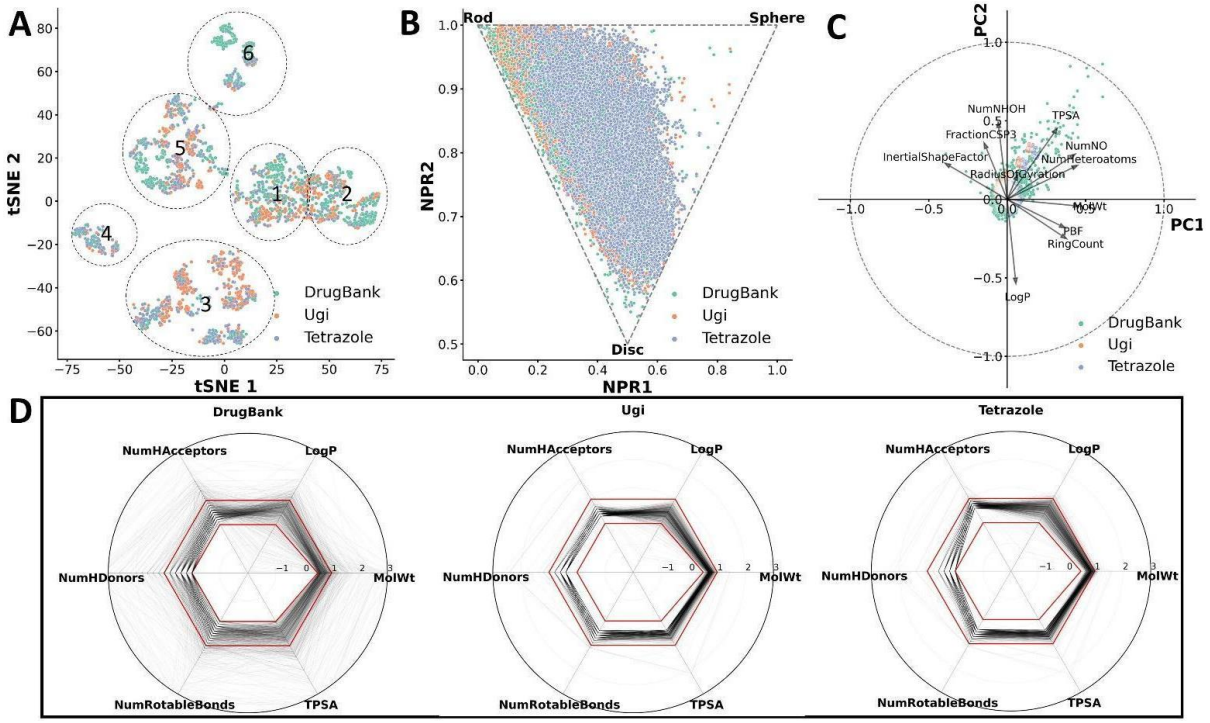


Figure 2. Characteristics of the generated libraries. (**A**) t-distributed stochastic neighbor embedding (tSNE) chart of structural diversity analysis for the compounds generated by Ugi three-component reaction (Ugi) or Ugi four-component tetrazole synthesis (Tetrazole). For comparison, a database containing FDA-approved molecules was included (DrugBank). (**B**) Normalized principal moments ratio (NPR) analysis. (**C**) Principal components analysis (PCA) for molecular and physicochemical descriptors. (**D**) Lipinski's rule of five (Ro5) analysis, which included molecular weight (MolWt), logarithmic partition coefficient (LogP), number of hydrogen bond donors (NumHDonors) and acceptors (NumHAcceptors), and the topological polar surface area (TPSA). A gray line represents each compound, and the density indicates the frequency of compounds.

A normalized principal moments ratio (NPR) analysis was conducted to assess the molecular shape distribution of compounds. The results showed that the minimum energy conformers of the compounds from the three libraries presented similar 3D shapes, predominantly rod- and disk-shaped, with only a few compounds displaying a spherical shape (Figure 2B). We also performed a principal components analysis (PCA) employing non-redundant molecular descriptors selected by their correlation (Figure S2). PCA showed that the compounds in CCC-generated libraries possess similar molecular descriptors and physicochemical properties to those of the DrugBank Database, displaying a dense accumulation of the compounds at the origin of the principal components (PCs), PC 1 and PC2. Topological polar surface area (TPSA) and logarithmic partition coefficient (LogP) were the primary descriptors, correlating positively with PC1 and PC2 and negatively with PC2, respectively (Figure 2C).

Finally, the extended Lipinski's rule of five (Ro5) analysis showed that most newly designed compounds comply with the physicochemical properties required for oral use [48]. Interestingly, the CCC-generated libraries displayed a more homogeneous distribution inside the extended Ro5 than the group of compounds in DrugBank (Figure 2D).

Virtual Screening

To identify new compounds with the theoretical ability to bind CD44, we generated expanded libraries containing 20 energetically favorable conformers for each compound within the CCC-generated libraries, retrieving 3,363,840 conformers. The expanded libraries were screened by alignment to the pharmacophore, followed by local optimization in CD44HAbd and visual inspection. We identified 864 conformers from 163 unique compounds that matched the selection criteria (Figure 3A). Only those molecules were employed for the subsequent experiments.

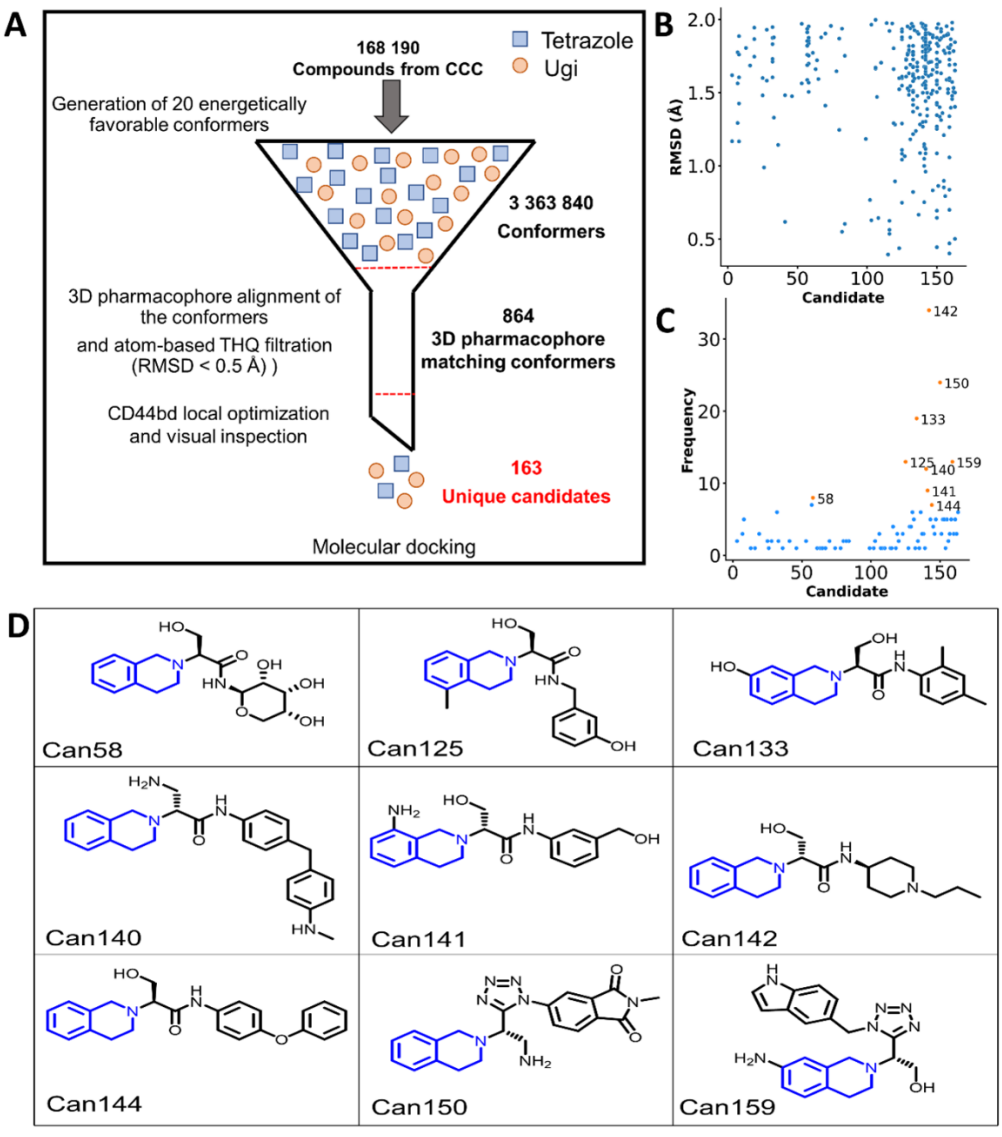


Figure 3. Virtual screening of THQ-containing molecules as potential CD44 antagonists. (**A**) The compounds within the CCC-generated libraries were sequentially filtered using the depicted strategy. (**B**) RMSD analysis comparing the THQ position in crystals vs. the poses obtained by docking of 163 unique compounds to human or mouse CD44HAbd. (**C**) Analysis of the frequency of poses with RMSD < 2 Å allowed the selection of nine candidates (orange dots). (**D**) Structure of the nine candidates selected by virtual screening with the THQ motif highlighted in blue.

The docking protocol employed for filtering was validated by docking the 21 THQ-containing molecules co-crystallized with mCD44HAbd into protein 5BZM. We observed that the crystal pose was more frequently reproduced in compounds with smaller substituents on the THQ motif (Figure S3A) Thus, additional analysis of candidates considered only the position of the THQ atoms.

The 163 candidates were docked against hCD44HAbd and mCD44HAbd for comparison. For each candidate/receptor pair, the docking scores of 25 poses were analyzed (Figure S4). The docking poses were compared to the coordinates of THQ crystalized on mCD44HAbd since none of the available human crystal structures contained THQ-derived compounds. We focused on the compounds showing docking poses matching the crystallized THQ atoms with RMSD < 2 Å (Figure 3B). We selected the compounds with the highest frequency of matching poses, ranging from 8/50 to 34/50 (Figure 3C). For the nine selected molecules (Figure 3D and Table S2), we assessed the THQ-binding site selectivity by docking the compounds into four/five additional pockets using three relevant forms of the receptor: apo-hCD44HAbd, HA-bound mCD44HAbd, and THQ-containing molecule mCD44HAbd. Except for the candidate (Can) 142, all molecules were predicted to bind the region of interest with better or similar affinity than other pockets (Figure S5). We then performed pose clustering analysis (Figure S6) to identify the best pose for molecular dynamics (MD) simulations.

Molecular Dynamics and Free Energy Calculation

We performed solvent explicit MD simulations to characterize the binding of the selected candidates to hCDHA44. The apo-hCD44HAbd was included as a control. Our analysis focused on the THQ binding site reported for mCD44HAbd (Figure 4A). By quantifying the water molecules displacement in the selected region, we identified that candidate Can58, Can133, Can141, Can142, and Can150 left the cavity during the simulation. Furthermore, Can58, Can142, and Can150 moved from the binding site at the early steps of the MD simulation. On the other hand, three compounds, (Can125, Can140, and Can159) remained on the binding site during the 100 ns of simulation and maintained a constant number of local water molecules

(Figure 4B). Interestingly, the global backbone RMSD analysis of the systems with those three candidates showed a different profile than the one generated by apo-CD44HAbd (Figure 4C) or by systems with other candidates (Figure S7). Additionally, a shift on some residues was observed on the root mean square fluctuation (RMSF) profile generated for Can125 (Figure 4D).

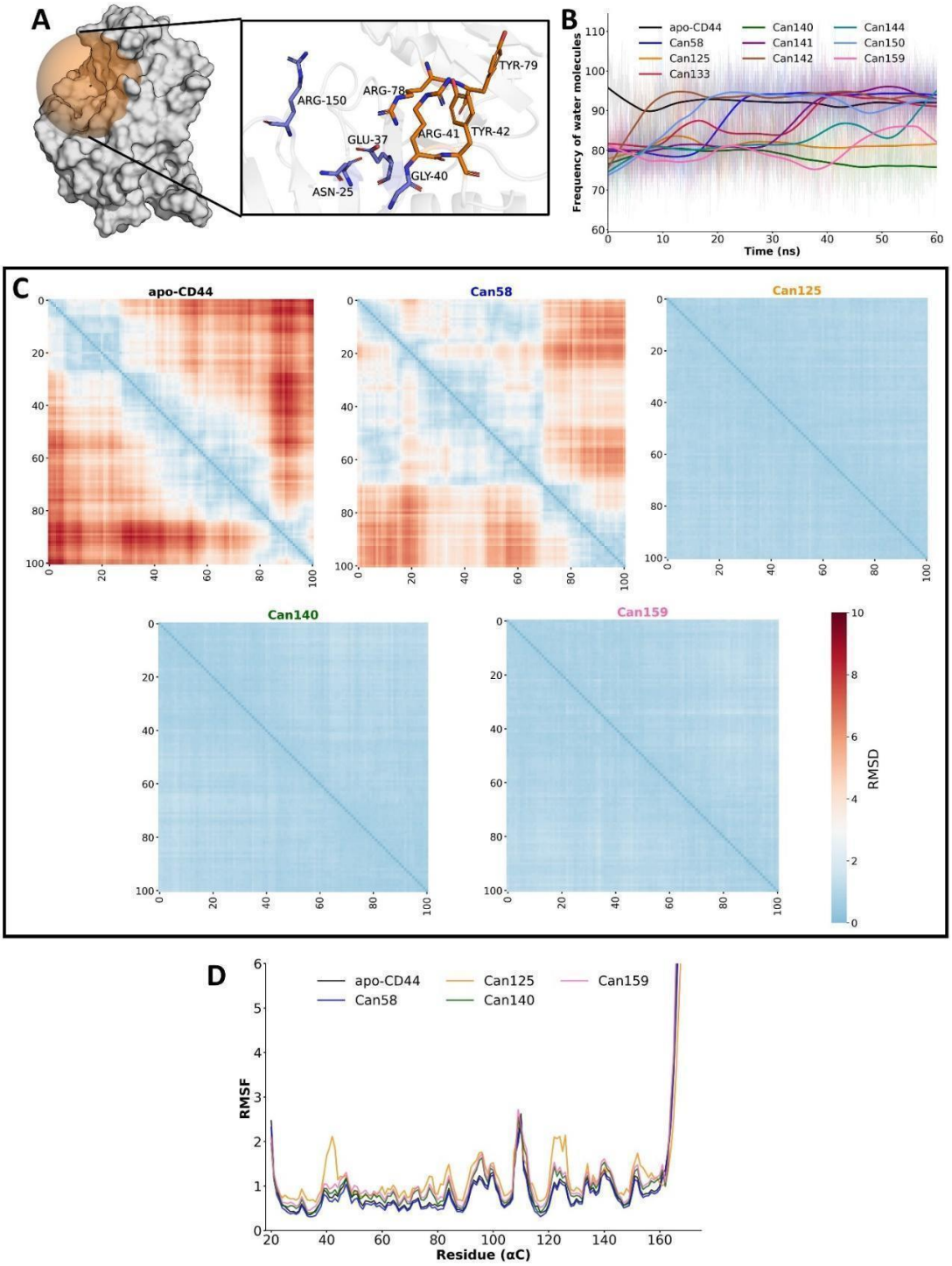


Figure 4. Molecular dynamics (MD) analysis identified Can 125 as a potential CD44 antagonist. (**A**) Spatial representation of the analyzed pocket in hCDHAbd. Inset shows the lateral chains of residues reported as essential for HA-binding (orange sticks) or those that mediate the interaction with the THQ-containing compounds employed for

pharmacophore modeling (blue sticks). (**B**) Water molecules displacement analysis for the nine candidates shown in **Figure 3**D. (**C**) Pairwise backbone RMSD matrix along 100 ns of MD simulation from systems including candidates with stable binding to hCD44HAbd. The unliganded protein (apo-CD44) and the system with Can58 are included for comparison. (**D**) Alpha-carbon RMSF analysis for the systems presented in **C**.

The frequencies of molecular interactions generated by Can125, Can140, and Can159 were studied during the whole simulation; Can58 was considered in the analysis for comparison (Figure S8). As expected from the water analysis, Can58 showed a low frequency of interactions in the THQ binding site, which included water bridges and hydrogen bonds with Arg150 and Glu75 (Figure 5A,B). Of the candidates studied by MD, Can125 formed the highest number of interactions, the most frequent were hydrogen bonds with Arg41 and Glu37, Van der Waals interactions with Arg150, Arg78, and Thr27, and water bridges with Arg78 and Glu37 (Figure 5A,C). Can140 showed Van der Waals interactions with residues Arg150, Asn25, and Phe74 predominantly, but it was also able to form water bridges with Arg78 and Glu37 (Figure 5A,D). Can159 showed a high frequency of π-cation, and hydrogen bonds interactions with Arg150, and water bridges with Arg78, and Arg150. Moreover, the Can159 also displayed the frequent formation of Van der Waals interactions with Asn25 (Figure 5A,E).

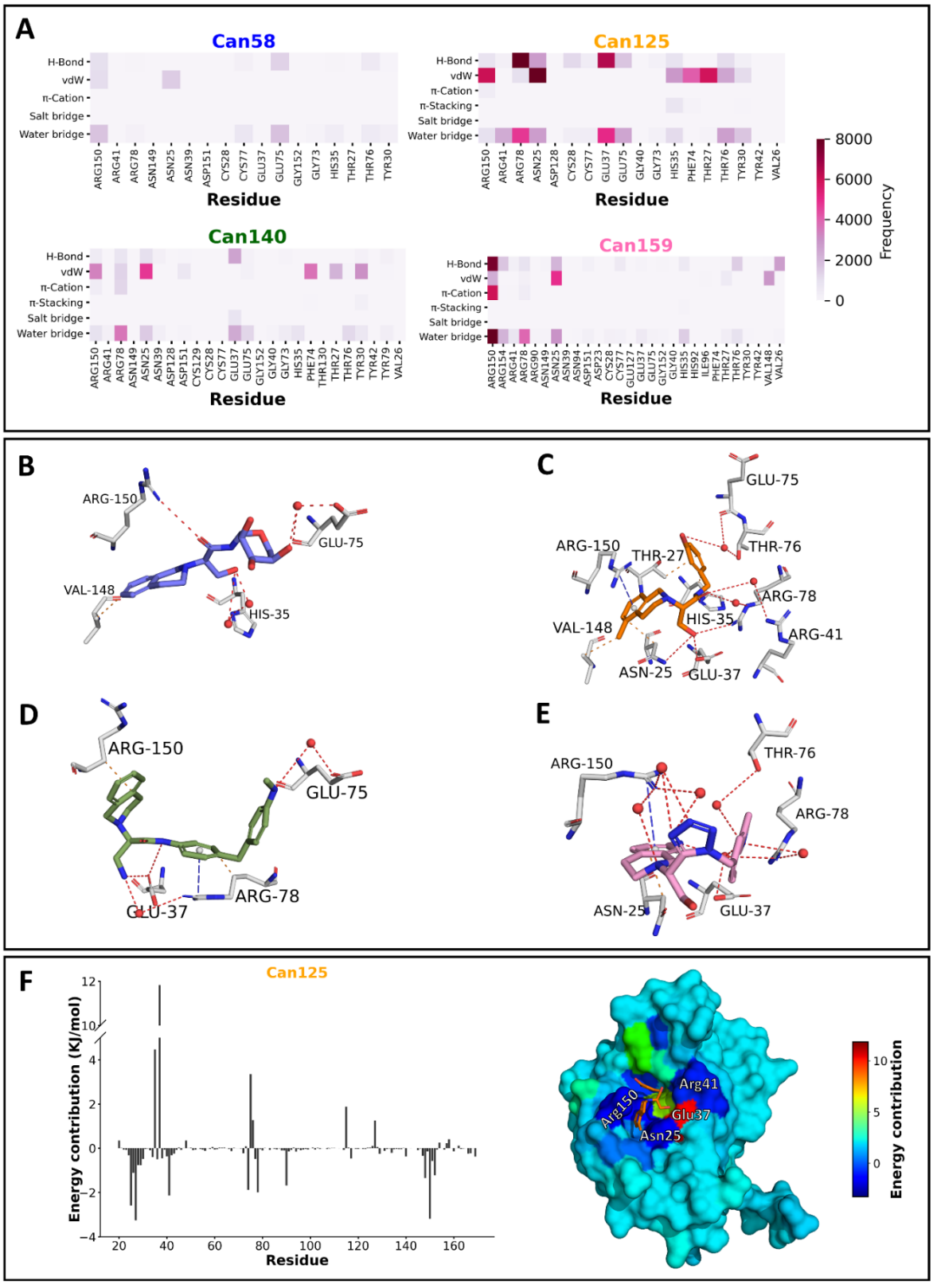


Figure 5. Identification and importance of residues mediating compound binding. (A) Types of interactions that support stable candidate/hCD44HAbd interactions and their frequency. Note that Can58, presented here for comparison, left the binding site during the simulation. (B–E) The 3D molecular interactions of a representative frame from MD simulation of the complexes between Can58 (B), Can125 (C), Can140 (D), and Can159 (E) with hCD44HAbd (gray sticks). (F)

Per-residue energy decomposition for Can125. Most positive values correspond to His35 and Glu37, and most negative values to Asn25, Thr27, Arg41, and Arg150.

Calculation of the binding free energy (Table 1) showed that Can140 has lower binding energy than Can125 and Can159, with the electrostatic energy as the component that contributes most to these differences. Nevertheless, Can140 showed large energy fluctuations during the simulations (Figure S9), suggesting a possible rearrangement of the binding pose during the experiment. In contrast, Can125 and Can159 showed stable energetic profiles (Figure S9). Per-residue energetic decomposition (Figure 5F and Figure S10) revealed that Can125 binding to hCD44HAbd is supported by energetically favorable interactions with residues essential for HA-binding (Arg41 and Arg78) or residues selected in the pharmacophore modeling (Arg150). Together, these observations suggest that Can125 is the most promising compound for biological evaluation.

Table 1. Average free energy calculated from MD simulations (mean ± standard deviation).

Compound	Van der Waals Energy	Electrostatic Energy	Polar Solvation Energy	SASA Energy	Binding Energy
Can58	−33.871 ± 35.190	−18.447 ± 29.906	47.592 ± 66.887	−4.667 ± 4.853	−9.393 ± 41.581
Can125	−116.512 ± 15.988	−52.356 ± 23.296	123.555 ± 27.579	−13.962 ± 1.252	−59.274 ± 17.744
Can140	−96.931 ± 36.947	−128.983 ± 66.843	106.522 ± 90.455	−12.078 ± 4.022	−131.470 ± 41.310
Can159	−99.395 ± 18.056	−34.928 ± 25.315	112.733 ± 40.928	−11.318 ± 1.938	−32.908 ± 17.750

Discussion

Due to the essential physiological and pathological roles of CD44, several crystal structures of its HAbd have been solved, either in the unligated form (apo) or in complex with HA. Our structural analysis of the crystals available at PDB corroborated the previously identified shifts in residues participating in HA-binding, including Arg41, Tyr42, Arg78, and Tyr79 in hCD44HAbd, and Arg45, Tyr46, Arg82, and Tyr83 in mCD44HAbd [22–24,49]. We found the shifts in Arg41 and Arg78 as particularly important for drug design because (i) the shift in Arg41 has been identified as a trigger for the conversion of high (active) to low (inactive) affinity conformations of CD44HAbd [22] and (ii) both residues are close to a pocket that binds to small molecules with the THQ motif and suffer conformation changes induced by ligand binding.

The THQ-binding pocket has been employed for the development of molecules that display a similar affinity for human or mouse CD44HAbd (ranging from 0.4 μM to 6.9 μM for hCD44HAbd, and 0.5 μM to 11.2 μM mCD44HAbd) [24]. Thus, we used 21 murine crystals containing THQ-based molecules for developing a pharmacophore that contained the key interactions mediating the binding of those compounds to mCD44HAbd and hypothesized that the model could be used in the identification of new antagonists for the human version of the receptor.

By using MCR-based CCC, we also generated libraries of easily synthesizable compounds that contain the THQ substructure. MCR are one-pot reactions in which two or more starting materials are used simultaneously; thus, most of the atoms from the initial building blocks are incorporated into the final product of the reaction [50,51]. The Ugi four-component tetrazole synthesis [43,44] and Ugi three-component reaction [45] are well described, easy to perform, and have been suggested as synthesis methods for diverse drug-like molecules [51]. Additionally, we selected building blocks that are commercially available at low-cost, which will allow compound synthesis and activity evaluation in further studies. Chemoinformatic characterization of the generated databases showed that the compounds are highly diverse but contain similar structural and physicochemical characteristics to those of marketable molecules. The DrugBank Database drugs frequently adopt rod- and disc-shapes [52] and comply with Lipinski's Ro5 [48,53]. The compounds within our CCC-generated libraries predominantly displayed these shapes, due to the high predominance of non-cyclic molecules [52,54], and have drug-like physicochemical characteristics. Molecules with the THQ substructure have been identified as nicotinic [55] or muscarinic [56] receptor antagonists, in addition to inhibitors of the angiotensin-converting enzyme [57]. Thus, the databases reported here may be useful starting points for identifying new compounds with those activities.

A robust exploration of the conformational space allowed the selection of 163 unique candidates that matched with the pharmacophoric model. To overcome the lack of structural information regarding the binding mode of THQ-containing molecules to hCD44HAbd, we investigated the binding mode of the candidates in hCD44HAbd and mCD44HAbd by a docking protocol that was able to reproduce the crystallographic binding mode of most co-crystallized molecules with an all-atoms RMSD threshold < 3Å. We identified nine candidates that reproduced the THQ crystallographic pose with an RMSD < 2Å and high frequency. A similar strategy, using a THQ-based pharmacophore for screening identified potential anticonvulsant compounds [58].

MD in explicit solvent further characterized the binding capability of the best nine candidates from virtual screening. We applied this method considering that the effects of solvation play a key role in forming molecular interactions in ligand–protein complexes. Thus, simulations employing explicit solvent allow the study of the most realistic and detailed level of physical chemistry of solvation [59]. We found that only three candidates (Can125, Can140, and Can159) remained bound to the THQ binding site during the entire MD simulation. In contrast, the Can142 left the THQ binding site at an early stage of the MD simulations; this observation might correlate with the fact that this candidate also showed higher docking scores for other pockets in CD44HAbd than the THQ-binding pocket (Figure S5).

Moreover, the candidates that remained bound during all simulations induced drastic decreases in the RMSD values of the hCD44HAbd backbone compared with those of the apo structure. The ligand-induced transition to a less flexible conformation of the protein can modulate its activation and improve both the compound's affinity and residence time [60,61]. Hence, a reduction in the target's conformational dynamics is a desirable characteristic of a drug-like molecule.

Although Can125, Can140, and Can159 displayed molecular interactions with residues involved in the HA-binding, including some reported as essential, only Can125 and Can159 reproduced the interactions predicted by the pharmacophore. Per-residue energetic decomposition corroborated that residues at the THQ-binding pocket support the binding of these two candidates to hCD44HAbd. However, only Can125 originated an RMSF profile that diverged from the unligated hCD44HAbd; specifically, it induced fluctuations in residues Arg41 and 120–126. The ligand-induced shift on Arg41 was different from the one identified on the crystal structures containing HA, which is considered essential for the transition from inactive to the active state in CD44 [22]. The changes in residues 120–126, which comprise a loop adjacent to the THQ-binding pocket, may participate in target constraint since they do not contribute to the ligand-binding energy. We hypothesize that the conformational changes induced by Can125, especially on Arg41, may impede the binding of the HA to hCD44HAbd.

On the other hand, Glu37 contributed negatively to the binding energy of Can125, which may be caused by the method employed for energy calculations. Calculations performed in implicit solvent offer a fast approach for binding energy assessment but are not yet well parameterized for complex problems that consider the presence of all solvent molecules [59]. Thus, our

calculations may be underestimating the energetic contribution of water bridges and hydrogen bonds formed between the side chain of Glu37 and the methyl alcohol of Can125. Moreover, it is also possible that the proximity between opposite hydrogen bond acceptors (the carboxylate of Glu37 and the oxygen in the acetamide group of Can125) represents an unfavorable energetic contribution. This finding represents an opportunity to improve the chemical features of potential antagonists of CD44 to be proposed in future studies.

We did not assess the possible effect of the best candidates in the binding of other CD44 ligands, as aggrecan, osteopontin, collagen, or CD74 [3,62,63], given the lack of corresponding structural data. However, we found that the THQ-binding site was able to allocate the best-ranked poses of Can125 and Can159 among all CD44HAbd pockets, suggesting a better affinity for this site over other regions of CD44HAbd. Thus, we speculate that these compounds may elicit competitive inhibition only for ligands with binding sites overlapping with that of HA, such as aggrecan [64]. On the other hand, we do not have evidence to propose that binding sites outside the CD44HAbd could be affected by the candidates, although the compounds restricted the conformational dynamics of CD44HAbd.

Conclusions

Our experiments demonstrate that the compounds Can125 [3-hydroxy-N-(3-hydroxybenzyl)-2-(5-methyl-3,4-dihydroisoquinolin-2(1H)-yl)propenamide] and Can159 [2-(1-((1H-indol-5-yl)methyl)-1H-tetrazol-5-yl)-2-(7-amino-3,4-dihydroisoquinolin-2(1H)-yl)ethan-1-ol] bind with high theoretical affinity to the murine and human structures of CD44HAbd and stabilize the conformational dynamics of the protein. Therefore, those compounds may elicit a blocking effect on HA-binding.

Supplementary Information

Table S1. mCD44HAbd crystal structures employed for 3D pharmacophore modeling

PDB	*Resolution*	*Year*	*Reference*
4MRE	1.58 Å	2014	[24]
4MRF	1.55 Å	2014	[24]
4MRG	1.69 Å	2014	[24]
4NP2	1.75 Å	2014	[24]
4NP3	1.61 Å	2014	[24]
5BZC	1.95 Å	2016	To be published
5BZE	1.31 Å	2016	To be published
5BZF	2.77 Å	2016	To be published
5BZG	2.19 Å	2016	To be published
5BZH	1.95 Å	2016	To be published
5BZI	1.32 Å	2016	To be published
5BZG	1.40 Å	2016	To be published
5BZK	1.40 Å	2016	To be published
5BZL	1.23 Å	2016	To be published
5BZM	1.25 Å	2016	To be published
5BZN	1.23 Å	2016	To be published
5BZO	1.22 Å	2016	To be published
5BZP	1.23 Å	2016	To be published
5BZQ	1.20 Å	2016	To be published
5BZR	1.15 Å	2016	To be published
5BZS	1.50 Å	2016	To be published
5BZT	1.25 Å	2016	To be published

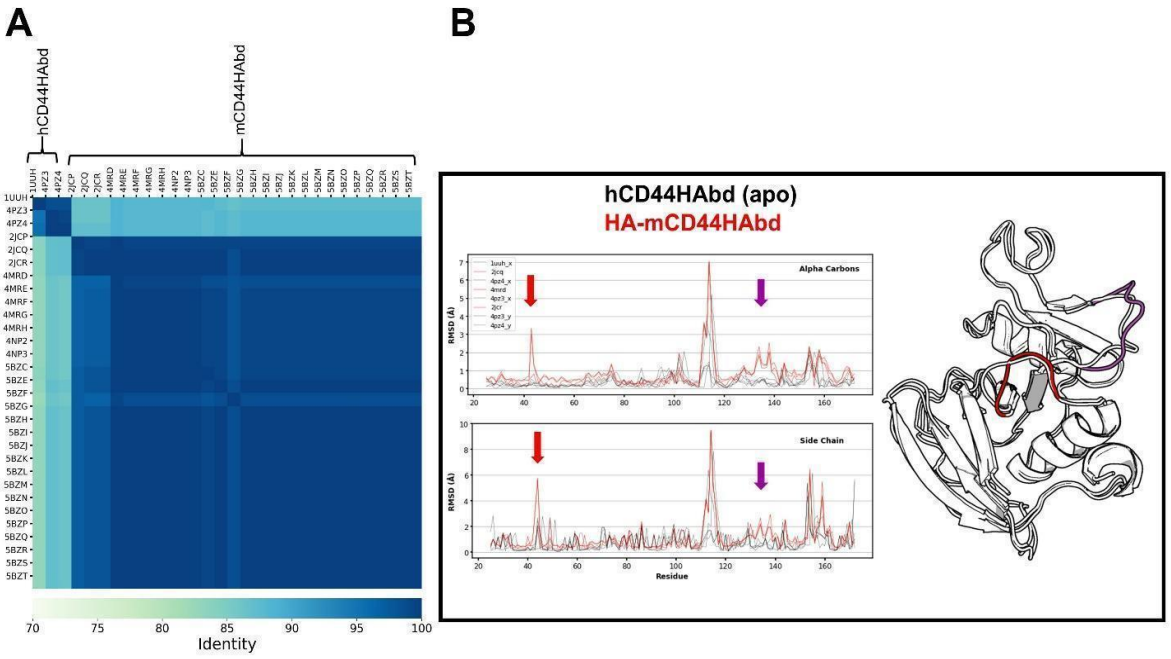


Figure S1. A) Comparison of the primary sequences of human (h) and mouse (m) CD44 HA-binding domain (CD44HAbd) available at PDB. B) RMSD calculated for selected ligand-free hCD44HAbd (apo) vs. HA-bound mCD44HAbd crystals (all structures co-crystallized with HA are murine). Two regions with significant structural changes associated with HA binding are indicated with arrows in graphs and colored in ribbons structure.

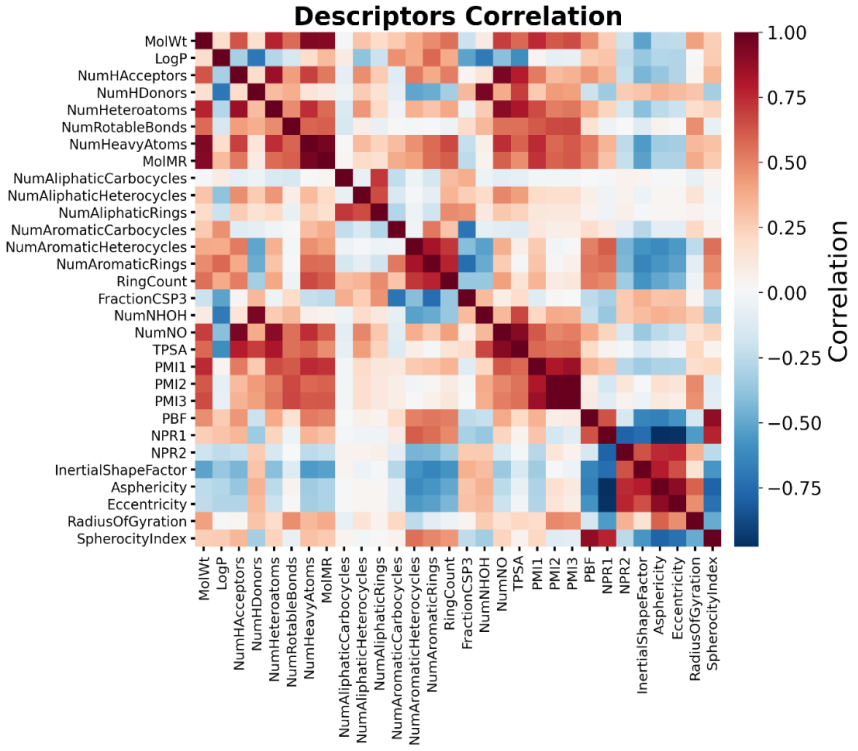


Figure S2. Correlation matrix of the molecular descriptors from the 168,190 compounds contained in our CCC-generated libraries. The matrix was employed for the selection of the non-redundant descriptors included in PCA presented in Figure 2.

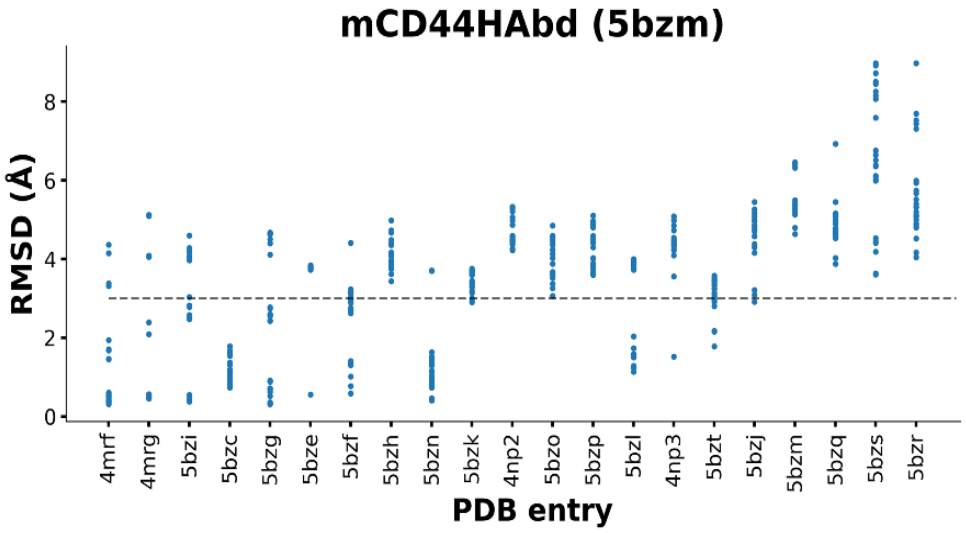


Figure S3. Validation of the docking protocol employed for candidate selection. RMSD of docking poses vs. crystal pose of 21 THQ-containing molecules. Molecules are identified by the PDB code in which they appear as ligands and are ordered from low (left) to high (right) molecular weight. Dashed line indicates the threshold considered for analysis.

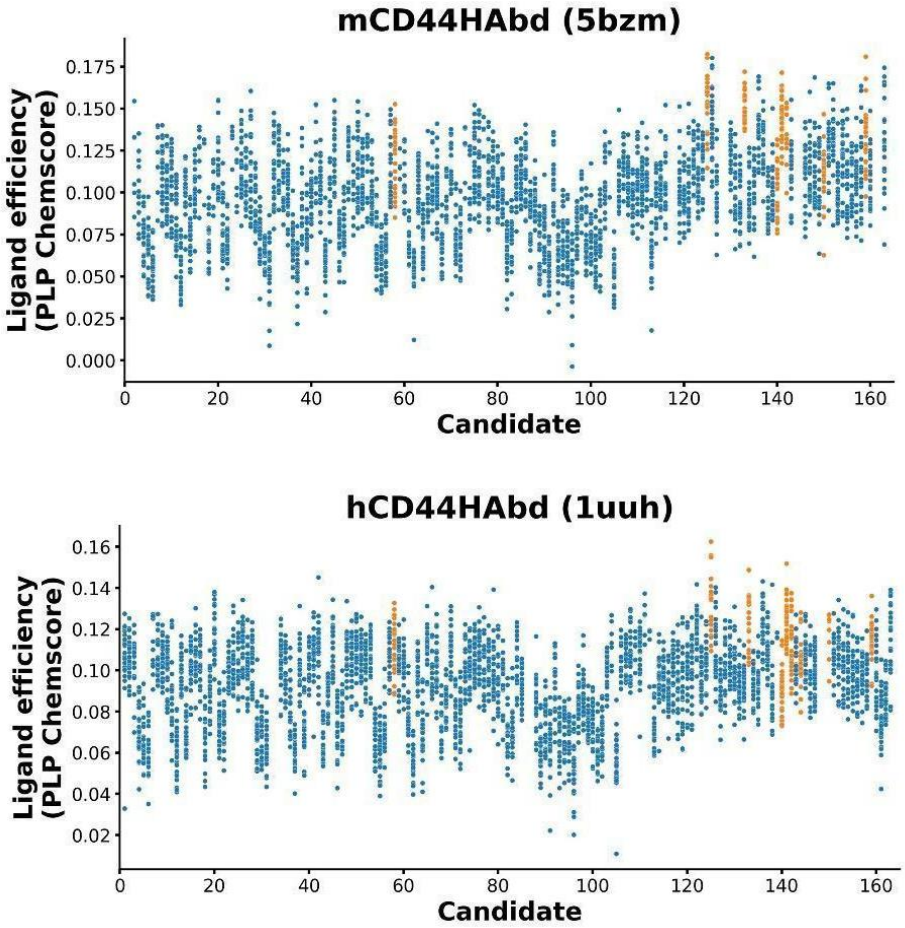
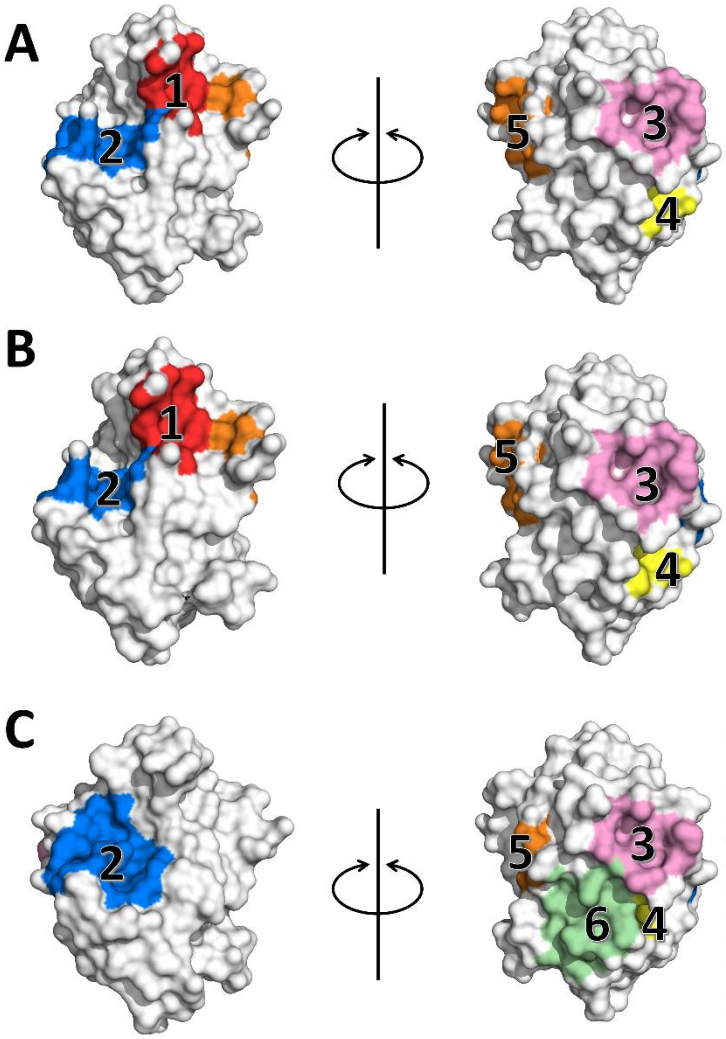


Figure S4. Ligand efficiency, calculated from docking scores, for the 163 unique candidates matching our 3D pharmacophore. The 25 best poses on mCD44 (A) and hCD44 (B) were analyzed. Values of the candidates selected for further analysis (see Figure 3) are labeled in orange.

Table S2. SMILES codes and formal names of the nine candidates presented in Figure 3D.

Code	SMILES	Name
Can58	OCC(C(N[C@@H](C1O)OCC(C1O)O)=O)N(C2)CCC3=C2C=CC=C3	2-(3,4-dihydroisoquinolin-2(1H)-yl)-3-hydroxy-N-((2R)-3,4,5-trihydroxytetrahydro-2H-pyran-2-yl)propanamide
Can125	CC1=CC=CC2=C1CCN(C2)C(C(NCC3=CC=CC(O)=C3)=O)CO	3-hydroxy-N-(3-hydroxybenzyl)-2-(5-methyl-3,4-dihydroisoquinolin-2(1H)-yl)propanamide
Can133	CC1=CC=C(C(C)=C1)NC(C(CO)N2CCC3=CC=C(C=C3C2)O)=O	N-(2,4-dimethylphenyl)-3-hydroxy-2-(7-hydroxy-3,4-dihydroisoquinolin-2(1H)-yl)propanamide
Can140	CNC1=CC=C(C=C1)CC2=CC=C(C=C2)NC(C(CN)N3CCC4=CC=CC=C4C3)=O	3-amino-2-(3,4-dihydroisoquinolin-2(1H)-yl)-N-(4-(4-(methylamino)benzyl)phenyl)propanamide
Can141	OCC1=CC(NC(C(N2CCC(C=CC=C3N)=C3C2)CO)=O)=CC=C1	2-(8-amino-3,4-dihydroisoquinolin-2(1H)-yl)-3-hydroxy-N-(3-(hydroxymethyl)phenyl)propanamide
Can142	CCCN1CCC(CC1)NC(C(CO)N2CCC3=CC=CC=C3C2)=O	2-(3,4-dihydroisoquinolin-2(1H)-yl)-3-hydroxy-N-(1-propylpiperidin-4-yl)propanamide
Can144	O=C(NC1=CC=C(C=C1)OC2=CC=CC=C2)C(CO)N3CCC4=CC=CC=C4C3	2-(3,4-dihydroisoquinolin-2(1H)-yl)-3-hydroxy-N-(4-phenoxyphenyl)propanamide
Can150	NCC(C1=NN=NN1C(C=C2C3=O)=CC=C2C(N3C)=O)N4CC5=CC=CC=C5CC4	5-(5-(2-amino-1-(3,4-dihydroisoquinolin-2(1H)-yl)ethyl)-1H-tetrazol-1-yl)-2-methylisoindoline-1,3-dione
Can159	NC1=CC=C2C(CN(CC2)C(C3=NN=NN3CC4=CC=C5NC=CC5=C4)CO)=C1	2-(1-((1H-indol-5-yl)methyl)-1H-tetrazol-5-yl)-2-(7-amino-3,4-dihydroisoquinolin-2(1H)-yl)ethan-1-ol



1uuh

Candidate	Score in pocket 2 (pocket rank)	Highest score in another pocket
Can58	60.25 (1)	49.64 [pocket 4]
Can125	52.75 (2)	57.34 [pocket 5]
Can133	56.18 (1)	54.82 [pocket 1]
Can140	59.02 (2)	65.81 [pocket 3]
Can141	53.66 (3)	58.85 [pocket 5]
Can142	47.61 (3)	59.20 [pocket 3]
Can144	57.10 (2)	66.13 [pocket 3]
Can150	58.26 (1)	55.86 [pocket 4]
Can159	66.75 (1)	56.95 [pocket 3]

2jcr

Candidate	Score in pocket 2 (pocket rank)	Highest score in another pocket
Can58	65.44 (1)	50.98 [pocket 5]
Can125	56.24 (2)	57.91 [pocket 3]
Can133	52.75 (1)	52.17 [pocket 3]
Can140	60.35 (2)	68.30 [pocket 3]
Can141	56.25 (2)	58.91 [pocket 5]
Can142	48.72 (3)	60.13 [pocket 3]
Can144	56.58 (2)	61.74 [pocket 3]
Can150	59.24 (1)	54.63 [pocket 3]
Can159	56.63 (2)	57.88 [pocket 3]

5bzm

Candidate	Score in pocket 2 (pocket rank)	Highest score in another pocket
Can58	59.03 (1)	53.92 [pocket 5]
Can125	62.72 (1)	58.92 [pocket 6]
Can133	56.41 (1)	52.81 [pocket 4]
Can140	65.85 (1)	62.71 [pocket 3]
Can141	60.56 (1)	57.47 [pocket 3]
Can142	52.33 (2)	59.67 [pocket 3]
Can144	63.00 (2)	65.01 [pocket 3]
Can150	62.37 (1)	52.67 [pocket 3]
Can159	65.52 (1)	58.72 [pocket 3]

Figure S5. Druggable pockets in apo-hCD44HAbd (**A**), HA-mCD44HAbd (**B**), and mCD44HAbd bound to a THQ-containing ligand (**C**), as predicted by Fpocket. The THQ-binding site corresponds to pocket 2 (blue). The comparison of binding scores between pockets (tables at the right part of figures) allowed assessment of the candidates' selectivity for the region of interest.

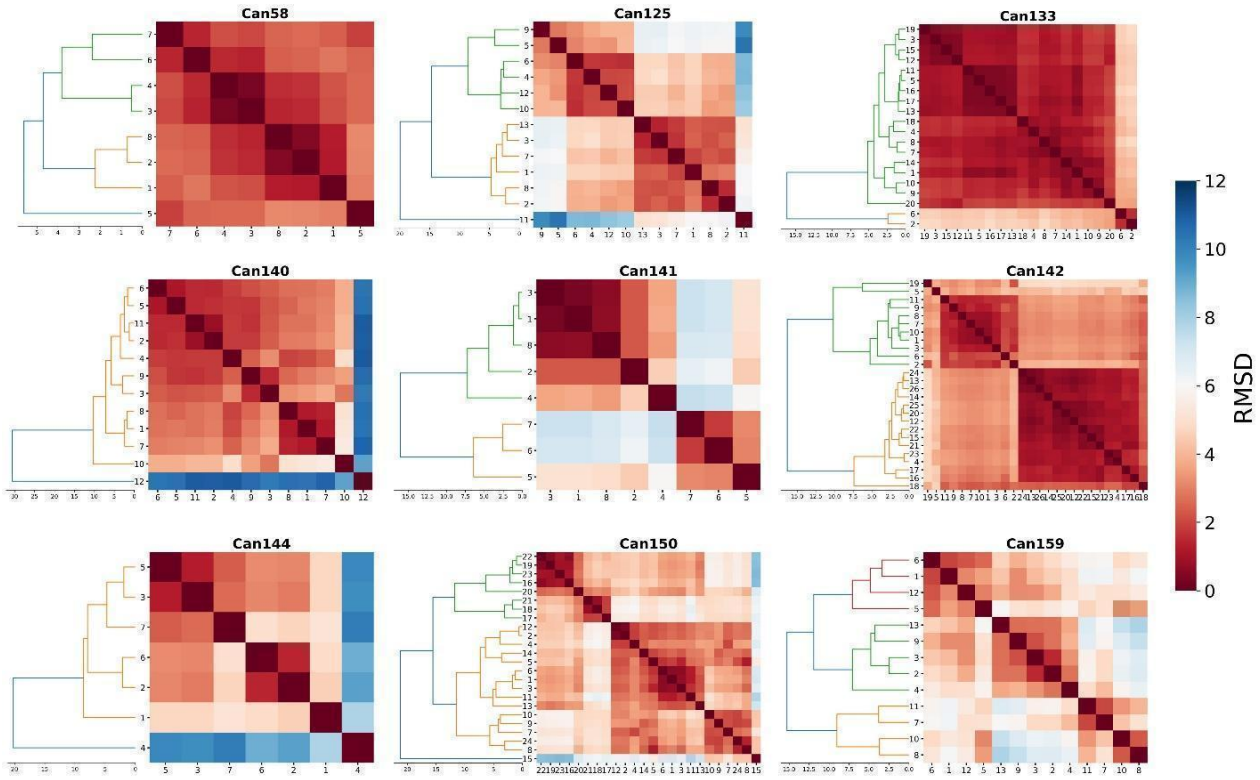


Figure S6. Docking pose clustering for the nine candidates with the higher frequency of poses resembling the crystallographic THQ pose. Only non-redundant poses (threshold RMSD >0.2 Å) are shown. These analyses allowed the identification of the most probable starting poses for MD simulations.

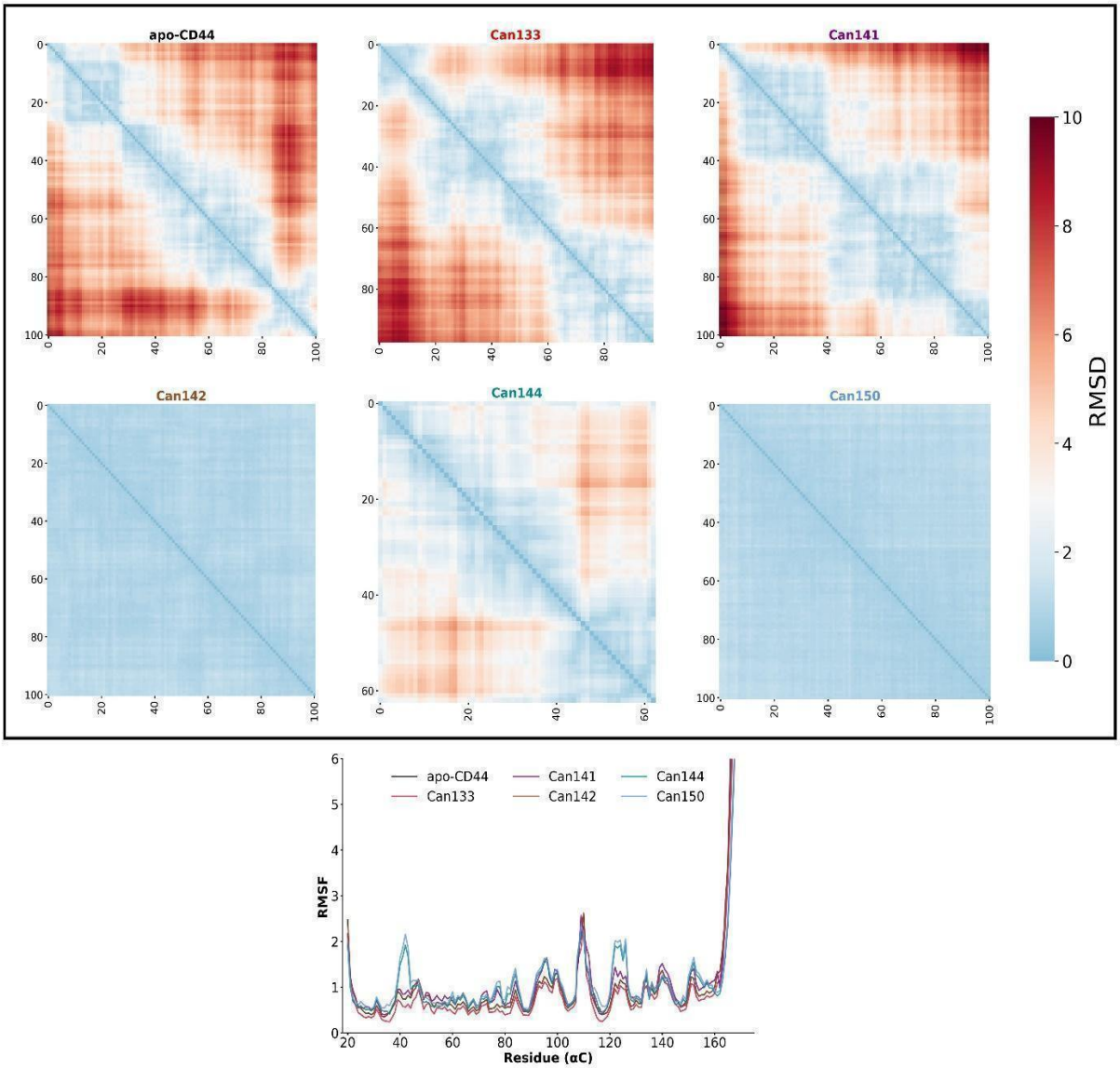


Figure S7. Global backbone RMSD matrix along 100 ns of MD simulation, and the corresponding alpha-carbon RMSF analysis, from systems with candidates (Can) with poor binding stability. The unliganded apoprotein (apo-CD44HAbd) is included for comparison.

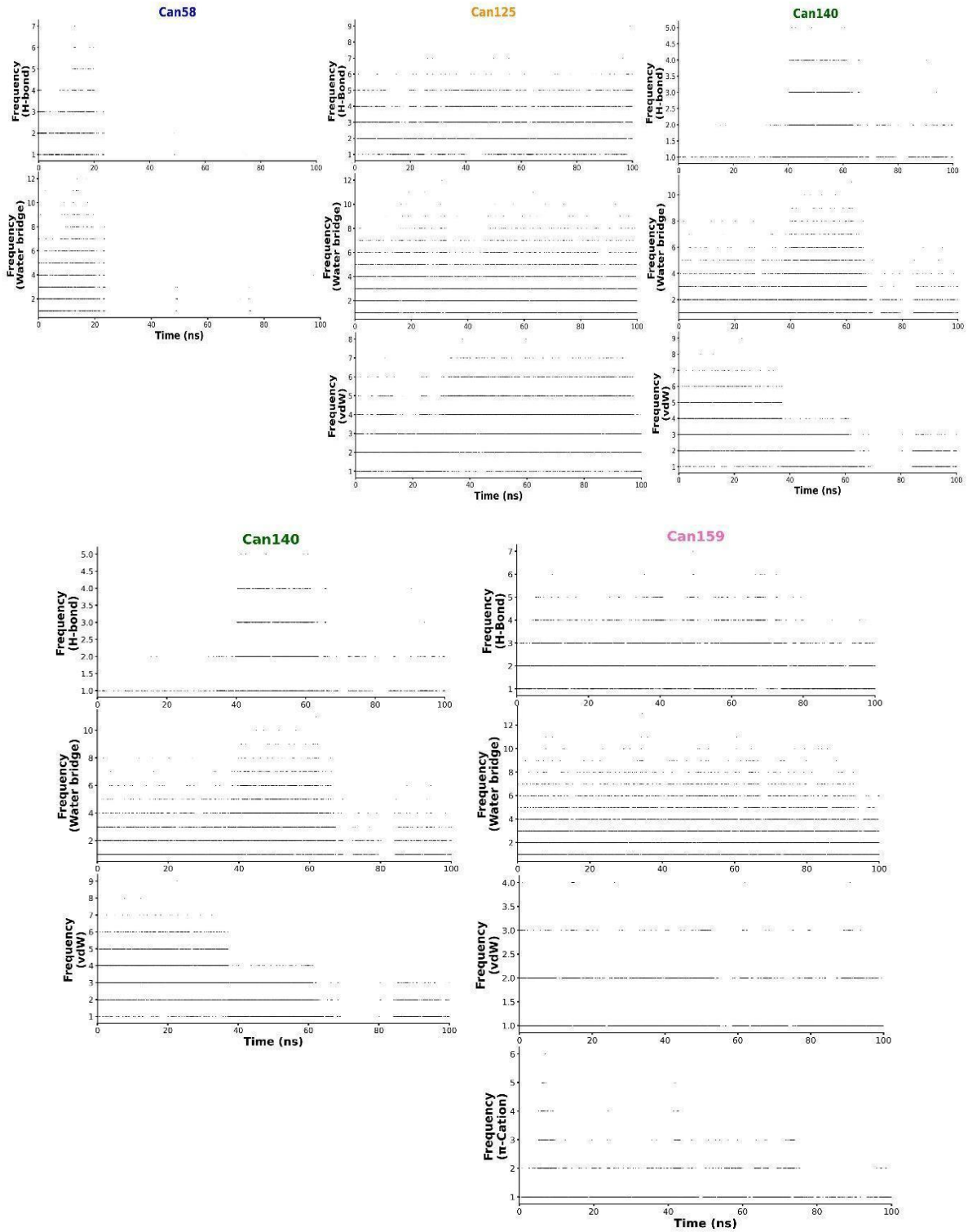


Figure S8. Frequency analysis of interactions employed for the generation of Figure 5A. Only interactions with frequency >2,000 are shown.

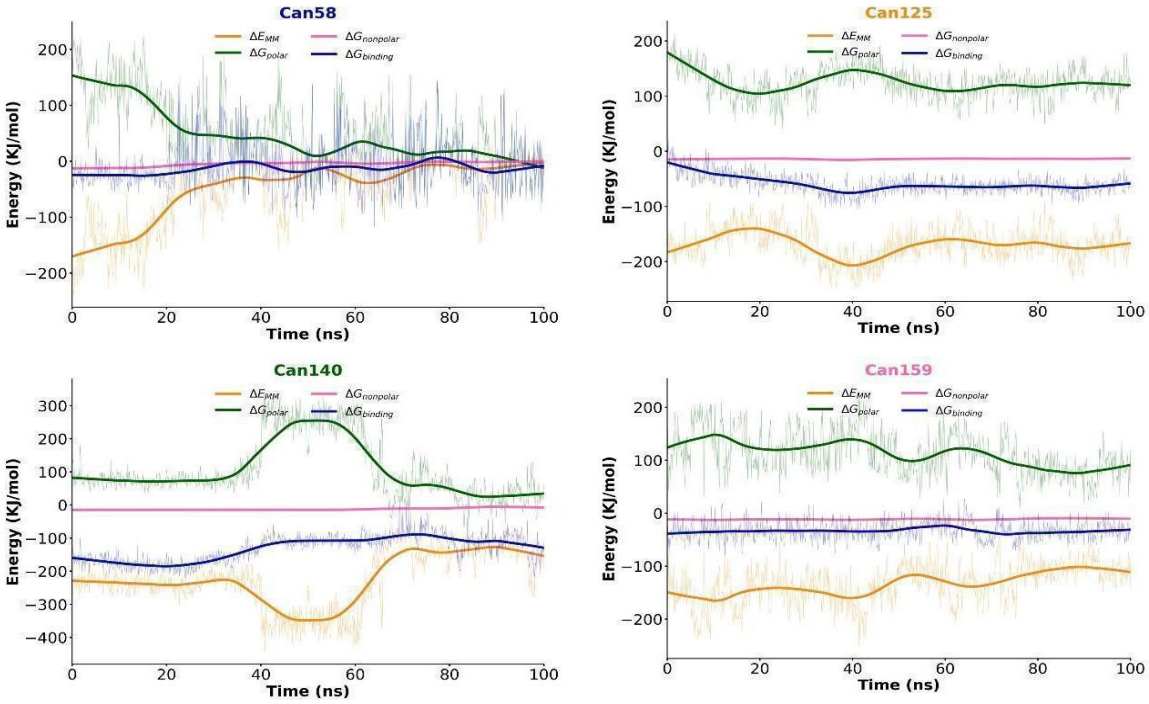


Figure S9. Energy calculations generated from MD simulations with candidates (Can) 125, 140, and 159. Can58 was employed as a negative control since it leaves the binding site during the simulation.

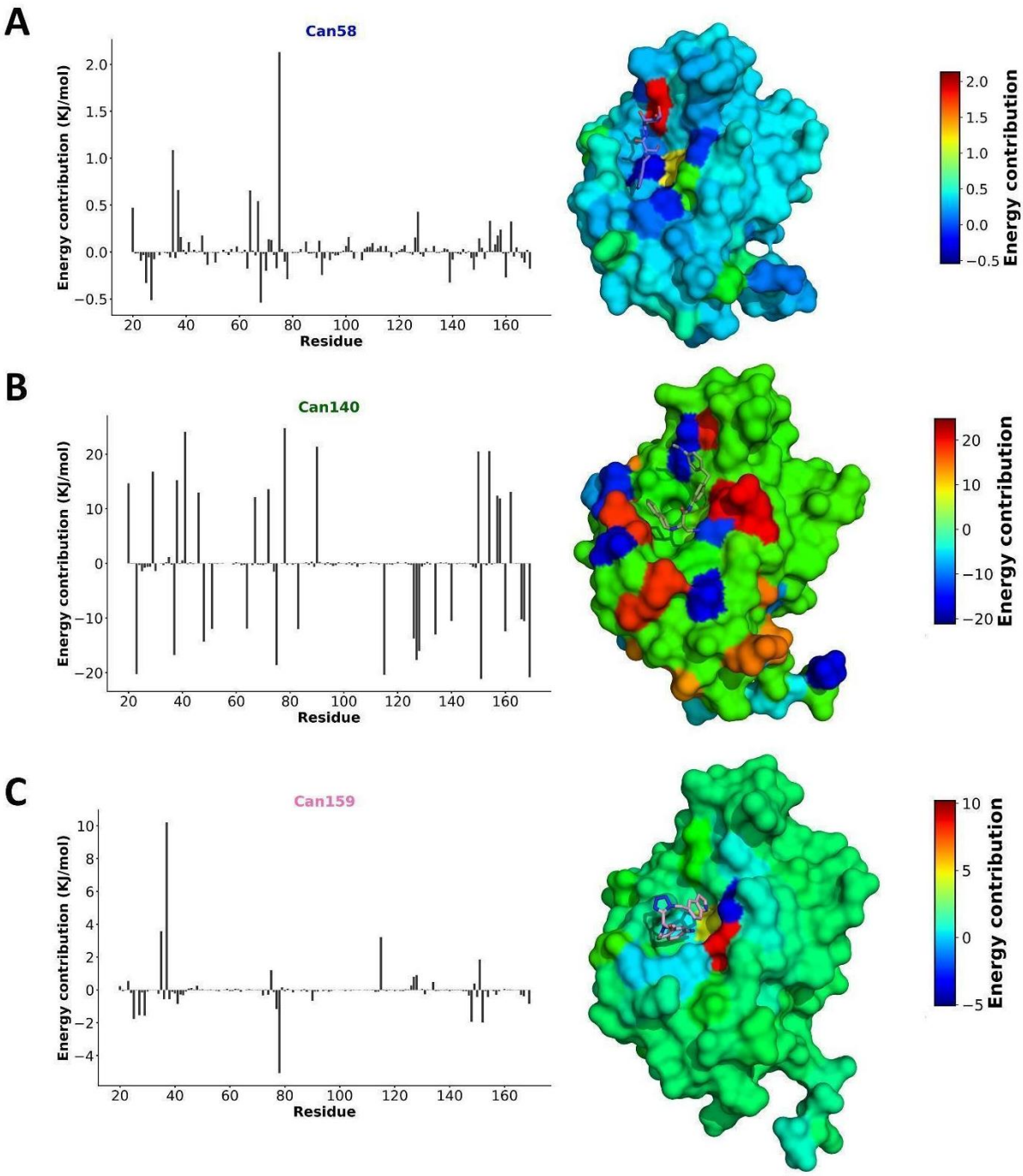


Figure S10. Per-residue energy decomposition for the MD-simulated binding of candidates (Can) 58 (**A**), 140 (**B**), and 159 (**C**) to hCD44HAbd.

References

1. Ponta H, Sherman L, Herrlich PA. CD44: From adhesion molecules to signalling regulators. Nat Rev Mol Cell Biol. 2003;4: 33–45. doi:10.1038/nrm1004

2. Zöller M. CD44: Can a cancer-initiating cell profit from an abundantly expressed molecule? Nat Rev Cancer. 2011;11: 254–267. doi:10.1038/nrc3023

3. Senbanjo LT, Chellaiah MA. CD44: A Multifunctional Cell Surface Adhesion Receptor Is a Regulator of Progression and Metastasis of Cancer Cells. Front Cell Dev Biol. 2017;5. doi:10.3389/fcell.2017.00018

4. Wu K, Xu H, Yuan X, Tian Y, Liu Y, Liu Q, et al. Enrichment of CD44 in basal-type breast cancer correlates with EMT, cancer stem cell gene profile, and prognosis. OTT. 2016; 431. doi:10.2147/OTT.S97192

5. Si D, Yin F, Peng J, Zhang G. High Expression of CD44 Predicts a Poor Prognosis in Glioblastomas. Cancer Manag Res. 2020;12: 769–775. doi:10.2147/CMAR.S233423

6. Wu G, Song X, Liu J, Li S, Gao W, Qiu M, et al. Expression of CD44 and the survival in glioma: a meta-analysis. Biosci Rep. 2020;40. doi:10.1042/BSR20200520

7. Chen J, Zhou J, Lu J, Xiong H, Shi X, Gong L. Significance of CD44 expression in head and neck cancer: a systemic review and meta-analysis. BMC Cancer. 2014;14: 15. doi:10.1186/1471-2407-14-15

8. Papadaki C, Manolakou S, Lagoudaki E, Pontikakis S, Ierodiakonou D, Vogiatzoglou K, et al. Correlation of PKM2 and CD44 Protein Expression with Poor Prognosis in Platinum-Treated Epithelial Ovarian Cancer: A Retrospective Study. Cancers (Basel). 2020;12. doi:10.3390/cancers12041013

9. Bourguignon LYW, Spevak CC, Wong G, Xia W, Gilad E. Hyaluronan-CD44 interaction with protein kinase Cϵ promotes oncogenic signaling by the stem cell marker nanog and the production of microRNA-21, leading to down-regulation of the tumor suppressor protein PDCD4, anti-apoptosis, and chemotherapy resistance. J Biol Chem. 2009;284: 26533–26546. doi:10.1074/jbc.M109.027466

10. Bourguignon LYW. Matrix Hyaluronan-CD44 Interaction Activates MicroRNA and LncRNA Signaling Associated With Chemoresistance, Invasion, and Tumor Progression. Front Oncol. 2019;9. doi:10.3389/fonc.2019.00492

11. Bourguignon LYW, Wong G, Shiina M. Up-regulation of Histone Methyltransferase, DOT1L, by Matrix Hyaluronan Promotes MicroRNA-10 Expression Leading to Tumor Cell Invasion and Chemoresistance in Cancer Stem Cells from Head and Neck Squamous Cell Carcinoma. J Biol Chem. 2016;291: 10571–10585. doi:10.1074/jbc.M115.700021

12. Hill A, McFarlane S, Mulligan K, Gillespie H, Draffin JE, Trimble A, et al. Cortactin underpins CD44-promoted invasion and adhesion of breast cancer cells to bone marrow endothelial cells. Oncogene. 2006;25: 6079–6091. doi:10.1038/sj.onc.1209628

13. Cieply B, Koontz C, Frisch SM. CD44S-hyaluronan interactions protect cells resulting from EMT against anoikis. Matrix Biol. 2015;48: 55–65. doi:10.1016/j.matbio.2015.04.010

14. Gudbergsson JM, Christensen E, Kostrikov S, Moos T, Duroux M, Kjær A, et al. Conventional Treatment of Glioblastoma Reveals Persistent CD44+ Subpopulations. Mol Neurobiol. 2020;57: 3943–3955. doi:10.1007/s12035-020-02004-2

15. Yu F, Yao H, Zhu P, Zhang X, Pan Q, Gong C, et al. let-7 Regulates Self Renewal and Tumorigenicity of Breast Cancer Cells. Cell. 2007;131: 1109–1123. doi:10.1016/j.cell.2007.10.054

16. Hong SP, Wen J, Bang S, Park S, Song SY. CD44-positive cells are responsible for gemcitabine resistance in pancreatic cancer cells. Int J Cancer. 2009;125: 2323–2331. doi:10.1002/ijc.24573

17. Dylla SJ, Beviglia L, Park I-K, Chartier C, Raval J, Ngan L, et al. Colorectal Cancer Stem Cells Are Enriched in Xenogeneic Tumors Following Chemotherapy. Gilliland DG, editor. PLoS ONE. 2008;3: e2428. doi:10.1371/journal.pone.0002428

18. Wang L, Huang X, Zheng X, Wang X, Li S, Zhang L, et al. Enrichment of Prostate Cancer Stem-Like Cells from Human Prostate Cancer Cell Lines by Culture in Serum-Free Medium and Chemoradiotherapy. Int J Biol Sci. 2013;9: 472–479. doi:10.7150/ijbs.5855

19. T H, S I, H N. Cancer stem-like cell marker CD44 promotes bone metastases by enhancing tumorigenicity, cell motility, and hyaluronan production. Cancer Res. 2013;73: 4112–4122. doi:10.1158/0008-5472.can-12-3801

20. Su Y-J, Lai H-M, Chang Y-W, Chen G-Y, Lee J-L. Direct reprogramming of stem cell properties in colon cancer cells by CD44. EMBO J. 2011;30: 3186–3199. doi:10.1038/emboj.2011.211

21. Gao Y, Foster R, Yang X, Feng Y, Shen JK, Mankin HJ, et al. Up-regulation of CD44 in the development of metastasis, recurrence and drug resistance of ovarian cancer. Oncotarget. 2015;6: 9313–9326. doi:10.18632/oncotarget.3220

22. Banerji S, Wright AJ, Noble M, Mahoney DJ, Campbell ID, Day AJ, et al. Structures of the Cd44-hyaluronan complex provide insight into a fundamental carbohydrate-protein interaction. Nat Struct Mol Biol. 2007;14: 234–239. doi:10.1038/nsmb1201

23. Bajorath J, Greenfield B, Munro SB, Day AJ, Aruffo A. Identification of CD44 residues important for hyaluronan binding and delineation of the binding site. J Biol Chem. 1998;273: 338–343. doi:10.1074/jbc.273.1.338

24. Liu LK, Finzel BC. Fragment-based identification of an inducible binding site on cell surface receptor CD44 for the design of protein-carbohydrate interaction inhibitors. J Med Chem. 2014;57: 2714–2725. doi:10.1021/jm5000276

25. Baggio C, Barile E, Di Sorbo G, Kipps TJ, Pellecchia M. The Cell Surface Receptor CD44: NMR-Based Characterization of Putative Ligands. ChemMedChem. 2016;11: 1097–1106. doi:10.1002/cmdc.201600039

26. Aguirre-Alvarado C, Segura-Cabrera A, Velázquez-Quesada I, Hernández-Esquivel MA, García-Pérez CA, Guerrero-Rodríguez SL, et al. Virtual screening-driven repositioning of etoposide as CD44 antagonist in breast cancer cells. Oncotarget. 2016;7: 23772–23784. doi:10.18632/oncotarget.8180

27. Pustuła M, Czub M, Łabuzek B, Surmiak E, Tomala M, Twarda-Clapa A, et al. NMR fragment-based screening for development of the CD44-binding small molecules. Bioorg Chem. 2019;82: 284–289. doi:10.1016/j.bioorg.2018.10.043

28. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, et al. UCSF Chimera--a visualization system for exploratory research and analysis. J Comput Chem. 2004;25: 1605–1612. doi:10.1002/jcc.20084

29. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics. 2009;25: 1422–1423. doi:10.1093/bioinformatics/btp163

30. Sunseri J, Koes DR. Pharmit: interactive exploration of chemical space. Nucleic Acids Res. 2016;44: W442–W448. doi:10.1093/nar/gkw287

31. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011;12: 2825–2830.

32. Gerber PR, Müller K. MAB, a generally applicable molecular force field for structure modelling in medicinal chemistry. J Comput Aided Mol Des. 1995. doi:10.1007/BF00124456

33. Gerber PR. Topological Pharmacophore Description of Chemical Structures using MAB-Force-Field-Derived Data and Corresponding Similarity Measures. In: Carbó-Dorca R, Gironés X, Mezey PG, editors. Fundamentals of Molecular Similarity. Boston, MA: Springer US; 2001. pp. 67–81. Available: https://doi.org/10.1007/978-1-4757-3273-3_5

34. Le Guilloux V, Schmidtke P, Tuffery P. Fpocket: An open source platform for ligand pocket detection. BMC Bioinformatics. 2009;10: 168. doi:10.1186/1471-2105-10-168

35. Jones G, Willett P, Glen RC, Leach AR, Taylor R. Development and validation of a genetic algorithm for flexible docking. J Mol Biol. 1997;267: 727–748. doi:10.1006/jmbi.1996.0897

36. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nature Methods. 2020;17: 261–272. doi:10.1038/s41592-019-0686-2

37. Abraham MJ, Murtola T, Schulz R, Páll S, Smith JC, Hess B, et al. GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. SoftwareX. 2015;1–2: 19–25. doi:10.1016/j.softx.2015.06.001

38. Jo S, Kim T, Iyer VG, Im W. CHARMM-GUI: A web-based graphical user interface for CHARMM. Journal of Computational Chemistry. 2008;29: 1859–1865. doi:10.1002/jcc.20945

39. Gowers RJ, Linke M, Barnoud J, Reddy TJE, Melo MN, Seyler SL, et al. MDAnalysis: A Python Package for the Rapid Analysis of Molecular Dynamics Simulations. Proceedings of the 15th Python in Science Conference. 2016; 98–105. doi:10.25080/Majora-629e541a-00e

40. Salentin S, Schreiber S, Haupt VJ, Adasme MF, Schroeder M. PLIP: fully automated protein–ligand interaction profiler. Nucleic Acids Res. 2015;43: W443–W447. doi:10.1093/nar/gkv315

41. Genheden S, Ryde U. The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. Expert Opin Drug Discov. 2015;10: 449–461. doi:10.1517/17460441.2015.1032936

42. Kumari R, Kumar R, Consortium OSDD, Lynn A. g_mmpbsa—A GROMACS Tool for High-Throughput MM-PBSA Calculations. American Chemical Society; 19 Jun 2014 [cited 26 Dec 2020]. doi:10.1021/ci500020m

43. Zhang J, Patil P, Kurpiewskab K, Kalinowska-T\luścikb J, Dömling A. Hydrazine in the Ugi Tetrazole Reaction Syn thesis. Synthesis (Stuttg). 2016;48: A-I. doi:10.1055/s-0035-1561353

44. Neochoritis CG, Zhao T, Dömling A. Tetrazoles via Multicomponent Reactions. Chem Rev. 2019;119: 1970–2042. doi:10.1021/acs.chemrev.8b00564

45. Tripolitsiotis NP, Thomaidi M, Neochoritis CG. The Ugi Three-Component Reaction; a Valuable Tool in Modern Organic Synthesis. European Journal of Organic Chemistry. 2020;2020: 6525–6554. doi:https://doi.org/10.1002/ejoc.202001157

46. Wishart DS, Knox C, Guo AC, Shrivastava S, Hassanali M, Stothard P, et al. DrugBank: a comprehensive resource for in silico drug discovery and exploration. Nucleic Acids Res. 2006;34: D668–D672. doi:10.1093/nar/gkj067

47. Polton DJ. Installation and Operational Experiences With MACCS (Molecular Access System). ONLINE REVIEW. 1982;6: 8.

48. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settingsq. Advanced Drug Delivery Reviews. 2001; 24.

49. Liu LK, Finzel B. High-resolution crystal structures of alternate forms of the human CD44 hyaluronan-binding domain reveal a site for protein interaction. Acta Crystallogr Sect Struct Biol Commun. 2014;70: 1155–1161. doi:10.1107/S2053230X14015532

50. Zarganes-Tzitzikas T, Chandgude AL, D??mling A. Multicomponent Reactions, Union of MCRs and beyond. Chem Rec. 2015;15: 981–996. doi:10.1002/tcr.201500201

51. Dömling A, Wang W, Wang K. Chemistry and biology of multicomponent reactions. Chem Rev. 2012;112: 3083–3135. doi:10.1021/cr100233r

52. Awale M, Reymond J-L. Web-based 3D-visualization of the DrugBank chemical space. Journal of Cheminformatics. 2016;8: 25. doi:10.1186/s13321-016-0138-2

53. Doak BC, Zheng J, Dobritzsch D, Kihlberg J. How Beyond Rule of 5 Drugs and Clinical Candidates Bind to Their Targets. J Med Chem. 2016;59: 2312–2327. doi:10.1021/acs.jmedchem.5b01286

54. Li J, Di Lorenzo V, Patil P, Ruiz-Moreno AJ, Kurpiewska K, Kalinowska-Tłuścik J, et al. Scaffolding-Induced Property Modulation of Chemical Space. ACS Comb Sci. 2020;22: 356–360. doi:10.1021/acscombsci.0c00072

55. Bowman WC. Neuromuscular block. Br J Pharmacol. 2006;147 Suppl 1: S277-286. doi:10.1038/sj.bjp.0706404

56. Doroshyenko O, Fuhr U. Clinical pharmacokinetics and pharmacodynamics of solifenacin. Clin Pharmacokinet. 2009;48: 281–302. doi:10.2165/00003088-200948050-00001

57. Asmar R, Sayegh F, Tracz W, Hlawaty M, Olszowska M, Jourde M, et al. Reversal of left ventricular hypertrophy with the ACE inhibitor moexipril in patients with essential hypertension. Acta Cardiol. 2002;57: 31–32.

58. De Luca L, Gitto R, Barreca ML, Caruso R, Quartarone S, Citraro R, et al. 3D Pharmacophore Models for 1,2,3,4-Tetrahydroisoquinoline Derivatives Acting as Anticonvulsant Agents. Arch Pharm Chem Life Sci. 2006;339: 388–400. doi:10.1002/ardp.200600022

59. Zhang LY, Gallicchio E, Friesner RA, Levy RM. Solvent models for protein–ligand binding: Comparison of implicit solvent poisson and surface generalized born models with explicit solvent simulations. Journal of Computational Chemistry. 2001;22: 591–607. doi:https://doi.org/10.1002/jcc.1031

60. Amaral M, Kokh DB, Bomke J, Wegener A, Buchstaller HP, Eggenweiler HM, et al. Protein conformational flexibility modulates kinetics and thermodynamics of drug binding. Nature Communications. 2017;8: 2276. doi:10.1038/s41467-017-02258-w

61. Ferina J, Daggett V. Visualizing Protein Folding and Unfolding. Journal of Molecular Biology. 2019;431: 1540–1564. doi:10.1016/j.jmb.2019.02.026

62. Naor D, Sionov RV, Ish-Shalom D. CD44: Structure, Function and Association with the Malignant Process. In: Vande Woude GF, Klein G, editors. Advances in Cancer Research. Academic Press; 1997. pp. 241–319. doi:10.1016/S0065-230X(08)60101-3

63. Shi X, Leng L, Wang T, Wang W, Du X, Li J, et al. CD44 Is the Signaling Component of the Macrophage Migration Inhibitory Factor-CD74 Receptor Complex. Immunity. 2006;25: 595–606. doi:10.1016/j.immuni.2006.08.020

64. Fujimoto T, Kawashima H, Tanaka T, Hirose M, Toyama-Sorimachi N, Matsuzawa Y, et al. CD44 binds a chondroitin sulfate proteoglycan, aggrecan. International Immunology. 2001;13: 359–366. doi:10.1093/intimm/13.3.359

Chapter 3

Benchmark of Generic Shapes for Macrocycles

Atilio Reyes Romero†, Angel Jonathan Ruiz-Moreno†, Matthew R. Groves, Marco Velasco-Velázquez, and Alexander Dömling.

This chapter has been published in J. Chem. Inf. Model. 2020, 60, 6298−6313.

† Equal contributors

Abstract

Macrocycles target proteins that are otherwise considered undruggable because of a lack of hydrophobic cavities and the presence of extended featureless surfaces. Increasing efforts by computational chemists have developed effective software to overcome the restrictions of torsional and conformational freedom that arise as a consequence of macrocyclization. Moloc is an efficient algorithm, with an emphasis on high interactivity, and has been constantly updated since 1986 by drug designers and crystallographers of the Roche biostructural community. In this work, we have benchmarked the shape-guided algorithm using a dataset of 208 macrocycles, carefully selected on the basis of structural complexity. We have quantified the accuracy, diversity, speed, exhaustiveness, and sampling efficiency in an automated fashion and we compared them with four commercial (Prime, MacroModel, molecular operating environment, and molecular dynamics) and four open-access (experimental-torsion distance geometry with additional “basic knowledge” alone and with Merck molecular force field minimization or universal force field minimization, Cambridge Crystallographic Data Centre conformer generator, and conformator) packages. With three-quarters of the database processed below the threshold of high ring accuracy, Moloc was identified as having the highest sampling efficiency and exhaustiveness without producing thousands of conformations, random ring splitting into two half-loops, and possibility to interactively produce globular or flat conformations with diversity similar to Prime, MacroModel, and molecular dynamics. The algorithm and the Python scripts for full automatization of these parameters are freely available for academic use.

Introduction

Macrocycles comprise a (hetero)cyclic core of at least 12 atoms, with molecular weight typically between 500 and 2000 Da. Ring sizes of 8−11 atoms and 3−7 atoms are classified as medium and small cycles. Although some naturally occurring rings contain up to 50 atoms, 14, 16-, and 18-membered rings occur at a higher frequency [1]. Generally, they encompass a large variety of chemical structures that originate from macrocyclization of simple building blocks, for example, cyclopeptide [2], cycloalkanes, and cyclodextrins [3], or as a result of de novo total synthesis or semisynthetic routes [4]. Among their clinical applications as drugs, macrocycles are used in oncology (temsirolimus [5,6] and epothilone B derivatives [7,8]), as antibiotics (vancomycin, macrolides, and rifampicin), immunology (sirolimus and zotarolimus), and in dermatology (pimecrolimus) [9]. Other applications of macrocycles are in supramolecular chemistry (crown ethers [10], cryptands, catenanes, rotaxanes [11], and calixarenes). Recently, macrocycles have received growing attention in medicinal chemistry [12–15] because of their unique ability to disrupt protein−protein interactions [16], improve metabolic stability [17], and improve cellular permeability by conformational restriction [18–21] resulting in a higher oral bioavailability compared to noncyclic congeners. Although macrocycles are outside of Lipinsk's rule of five, these molecules are able to bind proteins that are otherwise considered challenging because of their lack of hydrophobic cavities where functional groups can be anchored [22,23]. It has been estimated that nearly 25% of the ring atoms can contribute to the contact area with the protein surface through nonpolar contacts. Nevertheless, both ring atoms and peripherals/substituents show the same probability to match a hotspot, suggesting that ligand-based drug design of macrocycles should take into account these two components in order to identify potent binders [24]. We have recently described multiple scaffolds of artificial macrocycles which are readily synthesizable using multicomponent reaction chemistry (MCR) [25–30] and investigated the structural basis of macrocycles targeting PD1−PDL1, p53−MDM2, and IL17A receptor interactions [30–33]. Thus, we are highly interested in computational tools to rapidly screen conformational space of large virtual macrocycle libraries as a filter to synthesize bioactive compounds. To date, several benchmarks demonstrated the feasibility of algorithms with the aim of producing macrocycle conformations with enough accuracy and uniqueness for common computer-aided drug design (CADD) strategies, such as docking and pharmacophore screening [34]. Some of these algorithms are based on distance geometry (DG) [35], inverse kinematics [36], genetic algorithms [37], molecular dynamics (MD) simulations implementing either low frequency modes [38] or normal-mode search steps plus energy minimization [39], and, most recently, Monte Carlo multiple minimum/mixed torsional/low mode [40]. Generally, these software

programs are distinguished on the basis of the strategy adopted to generate conformations, systematic or stochastic. For example, molecular operating environment (MOE), MacroModel (MM), Cambridge Crystallographic Data Centre (CCDC) conformer generator, and experimental-torsion DG with additional "basic knowledge" (ETKDG) belong to the stochastic search category. Nevertheless, a major issue with these techniques is the generation of large numbers of representative conformers. On the other hand, a problem related to systematic search methods is the constrained flexibility of the ring, which is often insufficiently sampled by rotating a single bond at a time. In contrast to noncyclic molecules, the change in a single bond rotation impacts all bonds in macrocycles. Developing methods for sampling macrocycle conformations or improving upon the currently existing methods without generating a large number of conformers is a key step in the exploration of macrocycles in drug discovery. The computational basis of finite Fourier transform of ring structures was developed in 1985 [41] and its first embedding within a specialized conformer generator for macrocycle conformational sampling was shown in the publication of Gerber and co-workers in 1988 [42]. Fourier representation of the atomic position for macrocycle sampling has the advantage of generating a number of conformations that depend solely on the number of atoms in the ring, with few other user defined parameters. In the original publication, the author assessed the extensive conformational space covered by the Moloc software by taking (E)-cyclodecene and s-cis/s-trans-caprolactam as two study cases, investigating the potential of their method in combination with NMR spectroscopy of a macrocyclic tetrapeptide as a third example. This resulted in an exhaustive set of low-energy conformations of macrocyclic systems generated automatically, reproducing the experimented observed conformations, including s-cis/s-trans-isomers and, finally, showing the potential application in modelling surface loops of proteins. Herein, we benchmark the Fourier-based algorithm using a database of 208 macrocycle crystal structures and compare the performances of Moloc with the commercial software Prime, MOE, MD, MM, and four open-access packages experimental-torsion DG with additional "basic knowledge" and with the minimization steps employing the Merck molecular force field (MMFF94s [43]) or the universal force field (UFF [44]), CCDC, and conformator. We systematically assess the accuracy, structural diversity, and speed. Moreover, concepts of exhaustiveness and sampling efficiency (SE) are introduced. The aim of our work is to identify software capable of producing diverse and accurate conformations for daily virtual screening (i.e., docking). Moreover, because significant conformational changes in total shape and volume guide the bioavailability of certain macrocycles [45], we believe that the application of this approach could efficiently identify generic shapes of membrane-permeating conformations. A summary of the different software and the theoretical principles behind their functionality are presented in Table 1.

Table 1. Free and commercial software for the conformation generation of macrocycles and their working principles

Methodology	Description	Usage
Moloc	Macrocycle shapes are characterized by a selection of harmonics which occur in an approximate Fourier representation of the atomic coordinates of the rings [42].	Free
Conformator	Incremental construction of conformers with torsional angle assignment and a new deterministic cluster algorithm [46].	Free
CCDC	Ring template libraries to describe ring geometries using based on the wealth of experimental data in CSD.	Commercial
ETKDG	Stochastic search method that utilizes dg together with knowledge derived from experimental crystal structures [47,48].	Free
MOE	Perturbation of an existing conformation along a md' trajectory using initial atomic velocities with kinetic energy focused on the low-frequency vibrational modes and energy minimization [38].	Commercial
Prime	Ring splitting to create to two half rings that are sampled independently and recombined [49].	Commercial
MD	Desmond from Schrödinger Suite 2014-4 chosen as a baseline method (Maestro Desmond interoperability tools; Schrödinger: New York, NY, 2014).	Commercial
MM	Brief md simulations followed by minimization and normal-mode search steps [39].	Commercial

Materials and methods

Dataset

For a direct comparison of Moloc with the commercial and free software, we used the dataset of 208 macrocycles of Sindhikara and co-workers [49], consisting of 130 crystal structures from the Cambridge crystallographic dataset [50], a subset of 60 structures from the Protein Data Bank (PDB [51]) selected by Watts and co-workers [39] accounting for diverse and challenging macrocyclic topologies (disulfide bridges, cross-linking amide bonds, and polycyclic rings, including cyclodextrins, polyglycines, cycloalkanes, and peptidic macrocycles) and 18 crystals from the Biologically Interesting Molecule Reference Dictionary (BIRD) dataset chosen on the basis of quality (low-temperature factors and/or resolution < 2.1 Å) and structural diversity. Further details about the full dataset composition can be found in the Supporting Information from Sindhikara and co-workers [49].

Preparation of the Input Structures

Nonbiased starting conformations were prepared by removing the initial crystallographic coordinates, the partial charges, and the explicit hydrogens. Processed structures were converted to isomeric SMILES, preserving the stereochemistry flags. The resulting SMILES codes were employed as input for conformational sampling by conformator, CCDC conformer generator, and ETKDG alone or in combination with the minimization steps employing the MMFF94s or UFF while for Moloc, a set of random three-dimensional (3D) structures were generated using Mol3d.

Software Tested and Parametrization

MOE, Prime, MM, and MD. Macrocycle sampling description and initial condition for Prime, MOE, MM, and MD can be found in the Methods section of Sindhikara and co-workers while the results of accuracy, diversity, and speed can be found in the Supporting Information [49].

Moloc is one of the first molecular modelling packages and has since been updated regularly in close collaboration with drug designers and crystallographers of the Roche biostructural community, encompassing numerous functions, such as conformational sampling, generation of 3D pharmacophores [52], similarity analysis, peptide and protein modeling, modules for X-ray data handling, and ligand-based drug design. The generic Fourier description of the shape of the ring atoms is based on the generation of a series of harmonics [42]. Radial and axial deviations are then applied until a generic shape is found. Once it is identified, the algorithm starts to build a number of conformations that is proportional to the ring size. Geometric deviations, such as bond length and angles, are fixed by minimizing against the MAB force field [53]. In order to launch a sampling job, the “Mcnf” module was run in batch with the parameters “w0” and “c3” to initiate randomization of input atomic 3D coordinates and preserve the stereochemistry of both E/Z bonds and sp3 carbon, respectively. The selection of unique conformations is based on energetic (0.1 kcal/mol) and structural -0.1 Å root mean square deviation (RMSD) for cross-rigid body superimposition- thresholds. The conformations were kept within an energetic threshold of 10 kcal/mol. A conformational job can be launched using either two-dimensional (2D) or 3D atomic coordinates that are generated using Mol3d. During the conformational sampling, inner symmetries and permutations are enumerated. The number of generic shapes used as a start guide for the generation of the conformers grows as the square of N(ln N) where N represents

the number of ring atoms. Finally, for assessment, the flexibility of the software, energetic threshold, and hydrogen bond term were activated for the conformational job.

Conformator is a conformer generator focused on the enhancement of molecular torsion based on the assessment of torsion angles from the rotatable bonds. Conformator consists of a torsion driver enhanced by an elaborate algorithm for the assignment of torsion angles to rotatable bonds and a new clustering component that efficiently compiles ensembles by taking advantage of lists of partially presorted conformers. The clustering algorithm minimizes the number of comparisons between pairs of conformers that are required to effectively derive individual RMSD thresholds for molecules and to compile the ensemble. For this purpose, conformator features two conformer generation modes, “fast” and “best”, where “best” and “fast” focuses on the accuracy or speed of conformer search to generate conformers with the lowest RMSD values against a reference, respectively. Both modes attempt to ensure chemically correct bond angles and lengths as well as the planarity of aromatic rings and conjugated systems. After conformer generation, conformator performs a local optimization employing the macrocyclic optimization score which includes several well-known components from common force fields and some components specific to the optimization of macrocycles [46]. For optimal comparison of the software, we selected the “best” feature for macrocycle conformational sampling using the isomeric SMILES codes described above and requesting one thousand conformers per entry.

CCDC Conformer Generator. Conformer generator from CCDC is a knowledge-based method that uses data derived from CSD libraries and heuristic rules. For instance, conformer generator uses rotamer libraries to characterize preferred rotatable bond geometries and ring template libraries to describe ring geometries. Conformations are sampled based on CSD-derived rotamer distributions and ring templates. A final diverse set of conformers, clustered according to conformer similarity, are returned. Each conformer is locally optimized in torsion space [48,54]. For this work, the input structures described previously were loaded into the CCDC conformer generator through the CSD Python application programming interface (API). Conformer generator runs a minimization using the Tripos force field prior to conformational sampling, for which one thousand conformers were requested for each entry.

ETKDG Alone and with Minimization. RDKIT is an opensource toolkit for cheminformatics, comprising a wide variety of analysis and synthesis tools including similarity search, fingerprint

calculations, 2D and 3D descriptor calculation, and conformer generation (https://www.rdkit.org/). Currently, RDKIT can generate conformers using DG and an improved new method called ETKDG. The ETKDG algorithm is based on DG including experimental torsion angle termed experimental-torsion DG (ETDG) and "basic knowledge" (ETKDG) of molecular terms, including linear triple bonds and planar aromatic rings. The ETKDG method has been demonstrated to be more accurate in reproducing crystal structure conformations than DG alone. In addition, this algorithm has been recently optimized by the implementation of knowledge-based terms, preference for the trans-amide configuration, and the control of eccentricity from 2D elliptical geometry [48]. Thereby, we decided to explore the ETKDG approach for macrocycle sampling. Because ETKDG conformational sampling lacks any step of minimization, we ran minimization steps after the ETKDG conformational job using MMFF94s or UFF over 400 iterations per conformer in order to explore the minimization effect on macrocycle conformational sampling. We used the Python API of RDKIT to generate one thousand conformers per entry from the input structures.

Comparison Parameters

Exhaustiveness

Not all the softwares compared exhaustively sampled conformational space; some of them were not able to generate the requested conformers for some of the input structures. For instance, no sampling was performed in the case conformator if the assignment of torsion angles to rotatable bonds failed for a specific structure because this is the flexibility determination method employed using such a software. Thus, we defined the term exhaustiveness as follows:

$$Exhaustiveness = \frac{Num.\,entries\ sampled}{Total\ entries}$$

Accordingly, exhaustiveness values equal to 1 indicate full sampling of all entries in the dataset. Correspondingly, decreased exhaustiveness values indicate fewer entries sampled.

Accuracy

Based on previous benchmarks of conformational sampling [38,39,46,49,55,56], we have used RMSD to quantify the accuracy of the conformers in reproducing the reported bioactive crystallographic coordinates. The lowest RMSD values between each conformational ensemble to the reference structure were calculated. Notably, we have quantified the ring atom accuracy (RMSDbackbone) in a separate manner from heavy atom accuracy (RMSDheavy atoms), as indicated in Figure 1. This is based on the recently described classification of contacts between the macrocycle and its target: side chain, peripheral functional groups, and backbone atoms to the receptor [24]. Typically, a relative RMSD cutoff below 2.0 Å is considered an acceptable accuracy [57]. However, because macrocycles are more complex and larger than small molecules, we considered an RMSDheavy atoms value up to 2.5 Å as reasonably accurate and RMSDheavy atoms values below 1.0 Å were treated as highly accurate. Finally, we used the cumulative function distribution (CDF) to evaluate the performance of the algorithm in sampling a specific percentage of the dataset below two RMSDbackbone threshold values 0.5 Å (highly accurate) and 1.0 Å (accurate).

Diversity and SE

In order to systematically assess the structural diversity of each conformational ensemble, we used torsional fingerprints (TFs) in a similar manner to Sindhikara and co-workers [49]. The unique conformers were identified using a torsional scan on multiple conformations of a truncated version of the molecule comprising only the macrocycle backbone. Correspondence between related molecules was assessed by atom mapping from a maximum common substructure analysis. Then, a comparison of the fingerprints between the conformers was calculated using the torsional fingerprint deviation (TFD) [58]. Conformers with unique fingerprints were identified and kept if TFD was nonzero. As a further descriptor for assessment of shape diversity, we used the span in the radius of gyration (RoG), which is defined as the difference between the highest and the lowest RoG conformers [59]. Aiming to establish a relation among the exhaustiveness and the capability of the software to generate unique conformers, we introduced the SE as:

$$Sampling\ efficiency = Exhaustiveness\left(\frac{Unique\ Torsional\ Fingerprints}{Num.\ Conformers}\right)$$

SE values equal to 1 mean that each conformer represents a unique conformation within taking in account the number of entries sampled, while values close to 0 indicate high redundancy among conformers and/or lower exhaustiveness.

Speed

Time efficiency for each software was quantified by calculating the difference between the start and end time for conformer generation per entry. Batch scripts were generated for calculation of the time consumption for Moloc and conformator. Because of the usage of Python API for RDKIT and CCDC conformer generator, a tailored Python script was implemented in order to calculate the time consumption for CCDC conformer generator, ETKDG, and its further minimizations steps (UFF or MMFF94s). Moloc, conformator, and ETKDG alone or with minimization and CCDC conformer generator were run in a machine utilizing a 4-core Intel Xeon 3500 CPU-processor, 12 GB RAM, and 25 GB of data storage in a 1 TB HDD. The speed of MOE, MM, Prime, and MD was retrieved from the Supporting Information of the Prime benchmark publication [49].

Statistical Analysis

Data representation was carried out using the Python library matplotlib 3.1.1. [60] Statistical comparison of data was computed using a nonparametric Krustal–Wallis H-test among study groups using the stats module of SciPy [61]. All the p-values of the pairwise comparisons among the software can be found in the Supporting Information.

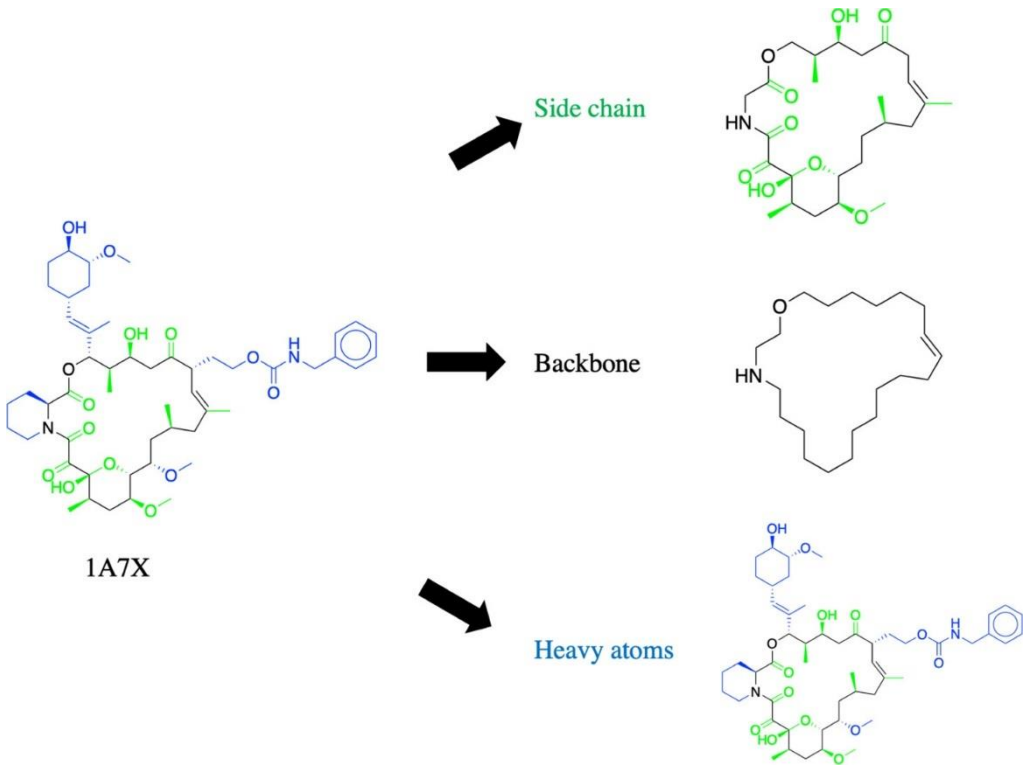


Figure 1. Example of separation of a 21-membered macrocycle into three atomic categories for the calculation of the RMSDbackbone and RMSDheavy atoms. Side chains, backbone, and heavy atoms are colored green, black, and blue, respectively.

Results

Exhaustiveness

According to our observations from conformational sampling of macrocycles employing different software, some methods were incapable of sampling all entries into the database. Conformator resulted in the least exhaustive sampling (190 out of 208 entries). Although the ETKDG algorithm was able to generate conformers for all input structures, the subsequent minimization step using UFF or MMFF94s force fields resulted in less exhaustiveness than the ETKDG algorithm alone (197 out of 208). All the remaining software tested (Moloc, CCDC conformer generator, and ETKDG) or previously reported (Prime, MOE, MM, and MD) was able to generate conformers for all input structures (Table 3).

Accuracy

Figure 2 indicates that all the software can generate conformers with reasonable accuracy (RMSDheavy atoms < 2.5 Å) and MM, MOE, and Prime generated conformers with median

RMSDheavy atoms values below a threshold of 1.0 Å with no statistical difference among the methods (Table S1). Among the six other software tested in this work, ETKDG algorithm plus MMFF94s minimization and Moloc were able to generate conformers with the lowest median RMSDheavy atoms value. However, in contrast to ETKDG plus MMFF94s minimization (0.9471), Moloc retained superior exhaustiveness (1), indicating that it can generate reasonably accurate conformers across a complex and diverse dataset of macrocycle molecules. No statistical difference was found among all open-source methods, including CCDC conformer generator. Finally, MD showed a median RMSDheavy atoms value slightly higher for the highly accurate threshold, and statistical difference versus all the remaining private and open-access methods. In RMSDbackbone and CDF analysis, Figure 2A shows that Prime, MM, MOE, and CCDC conformer generator produced the highest accurate conformers (RMSDbackbone < 0.5 Å) with no statistical difference among these four methods (Table S2), returning a fraction of entries sampled for each method of 0.63, 0.67, 0.58, and 0.46, respectively (Figure 2B and Table 2).

Table 2. Fraction of entries sampled below the two $RMSD_{backbone}$ thresholds chosen as highly accurate (<0.5 Å) and accurate (<1.0 Å).

Method	<0.5 Å	<1.0 Å
Prime	0.63	0.9
MM	0.67	0.9
MOE	0.58	0.8
MD	0.4	0.79
Moloc	0.31	0.79
Conformator	0.26	0.68
CCDC	0.46	0.65
ETKDG	0.19	0.72
MMFF94s	0.27	0.78
UFF	0.17	0.7

In addition, our data indicate that all the remaining methods generated conformers below 1.0 Å. No statistical difference was observed among MD, Moloc, and ETKDG with MMFF94s, whose fraction of sampled entries was, respectively, 0.79 for the first two and 0.78. Such results indicate similar accuracy among these methods to reproduce the reference macrocycle backbone structure. Similarly, no statistical difference was found between Moloc and MMFF94s and both produced a similar fraction of entries sampled above the threshold (Moloc: 0.77, MMFF94s: 0.79).

Finally, comparison between conformator, ETKDG, and ETKDG plus UFF minimization did not show any statistical differences. A statistical difference was found when comparing conformator, ETKDG, and ETKDG plus UFF minimization versus Moloc or ETKDG plus MMFF94s minimization with a fraction of entries sampled being 0.68 for conformator, 0.72 for ETKDG, and 0.70 for ETKDG plus UFF minimization steps. However, among these last groups of methods, ETKDG is the most exhaustive followed by ETKDG plus UFF minimization and conformator.

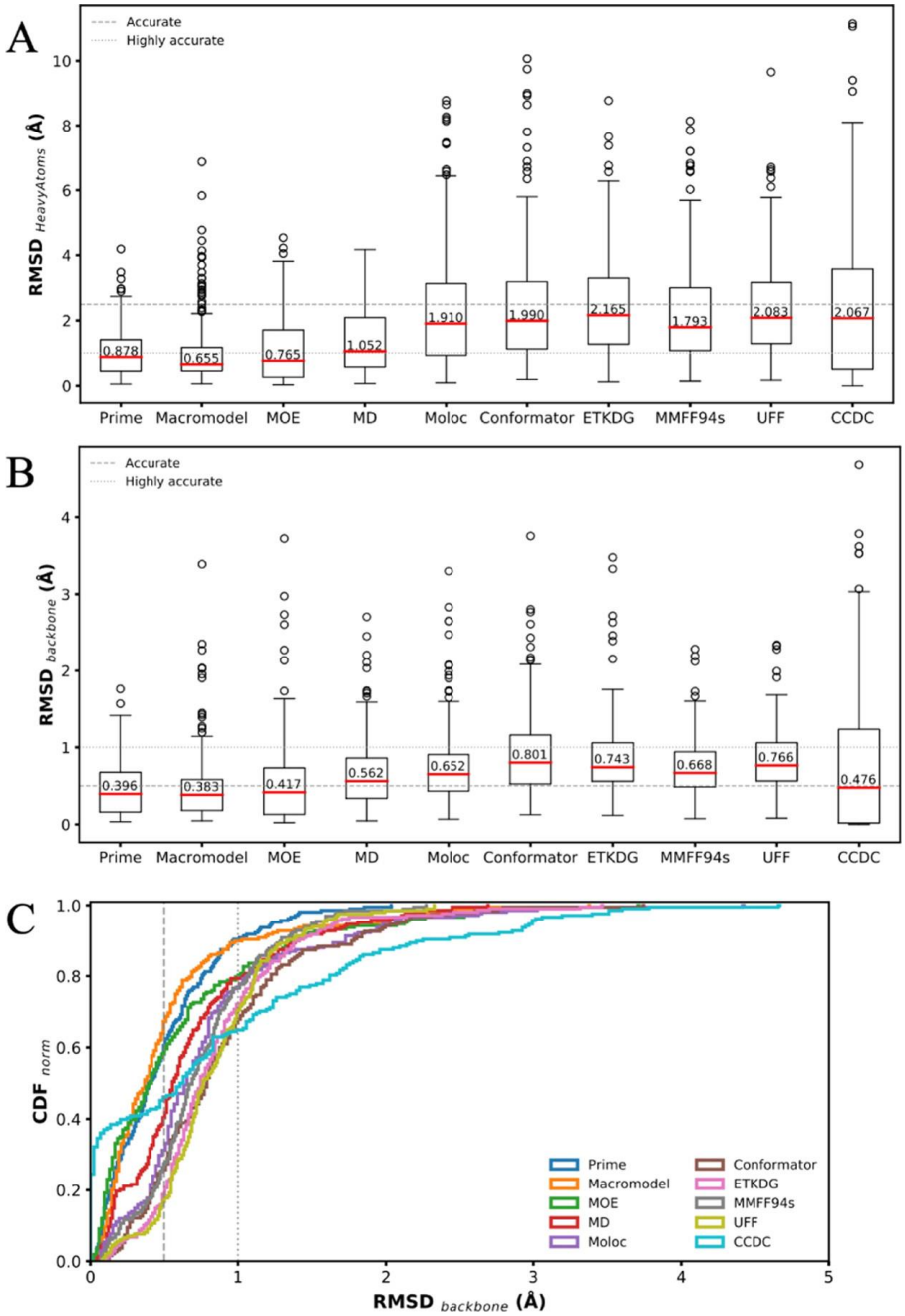


Figure 2. Crystal structure accuracies for each method displayed as (A) RMSDheavy atoms and (B) RMSDbackbone, respectively. (C) Normalized cumulative distribution function (CDFnorm). The accuracy threshold values, median, and outliers are presented as gray dots, red lines, and black contoured circles, respectively.

Diversity and SE

Although all software was challenged with a one thousand conformers per entry request, not all of them succeeded in accomplishing the task, either retrieving fewer conformers per entry or unable to sample some, resulting in poor exhaustiveness. Among the methods studied, only MD and ETKDG succeeded in generating all conformers requested. Nevertheless, we compared the TFs of the conformers for each method in order to assess the number of unique conformers generated and, furthermore, we employed the exhaustiveness value to calculate the SE of each software. We identified Moloc and ETKDG followed by ETKDG plus minimization with either MMFF94s or UFF as the most efficient methods to perform conformational search of macrocycles (Table 3).

Table 3. Summary table of the exhaustiveness and sampling efficiency, number of conformers, and torsional fingerprints.

Method	Exhaustiveness	Unique Torsional Fingerprints (median)	Number of conformers (median)	Sampling efficiency
Prime	208/208 =1	707	932	0.7586
MM	208/208 =1	100	300	0.3333
MOE	208/208 = 1	48	76	0.6316
MD	208/208 = 1	59	1000	0.059
Moloc	208/208 = 1	67	67	1
Conformator	190/208 = 0.9135	246	338	0.6648
ETKDG	208/208 = 1	1000	1000	1
MMFF94s	197/208 = 0.9471	998	998	0.9471
UFF	197/208 = 0.9471	535	535	0.9471
CCDC	208/208 = 1	6	8	0.75

On the contrary, although MD showed an exhaustiveness value of 1, it is also a highly redundant method generating only a median of 59 unique conformers across 1000 conformers retrieved, obtaining the lowest SE value (0.059) among all reported methods. In a similar fashion to MD, MM showed a low SE. Despite being a highly exhaustive methodology, the relation between the number of conformers generated and their uniqueness results in an SE of 0.333. Thus, Moloc and ETKDG are three times more efficient in macrocycle conformation sampling than MD. However, Prime (exhaustiveness: 1) was able to produce a median of 707 unique conformers for a median of 932 conformers, resulting in an SE of 0.7586. A similar behavior was observed for MOE, which obtained exhaustiveness equal to 1 and an SE of 0.6316. CCDC conformer generator showed an

SE of 0.7500 with the lowest number of unique conformers generated (Figure 3A) across all the software studied.

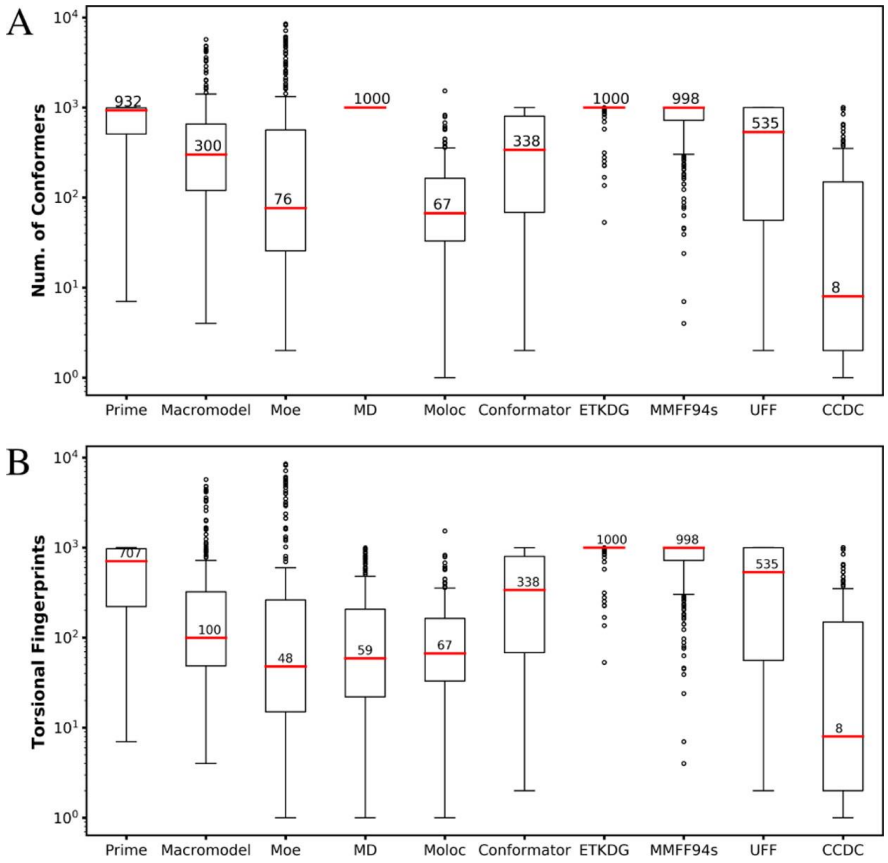


Figure 3. Panel showing (A) box plot of number of the conformers and (B) TFs for each method. Graphical description of median and outliers is the same as in Figure 2.

Figure 4A compares the results obtained from the span of RoG as a parameter to study the 3D conformational diversity of the conformers moving from a globular to a flat-shaped conformation (Figure 4B). Our data indicate that ETKDG algorithm plus MMFF94s minimization (1.13 Å) achieved the highest span in RoG with no statistical difference with Prime (1.02 Å) and ETKDG with UFF minimization (1.08 Å) (Table S4). On the other hand, the conformations produced by Moloc (0.86 Å) were proven to be statistically similar to MM (0.93 Å), MOE (0.74 Å), MD (0.85 Å), conformator (0.87 Å), and ETKDG alone without minimization (0.82 Å). Finally, with a span in RoG of 0.15 Å, the conformers produced by CCDC conformer generator were identified as having the lowest diversity among all the software tested.

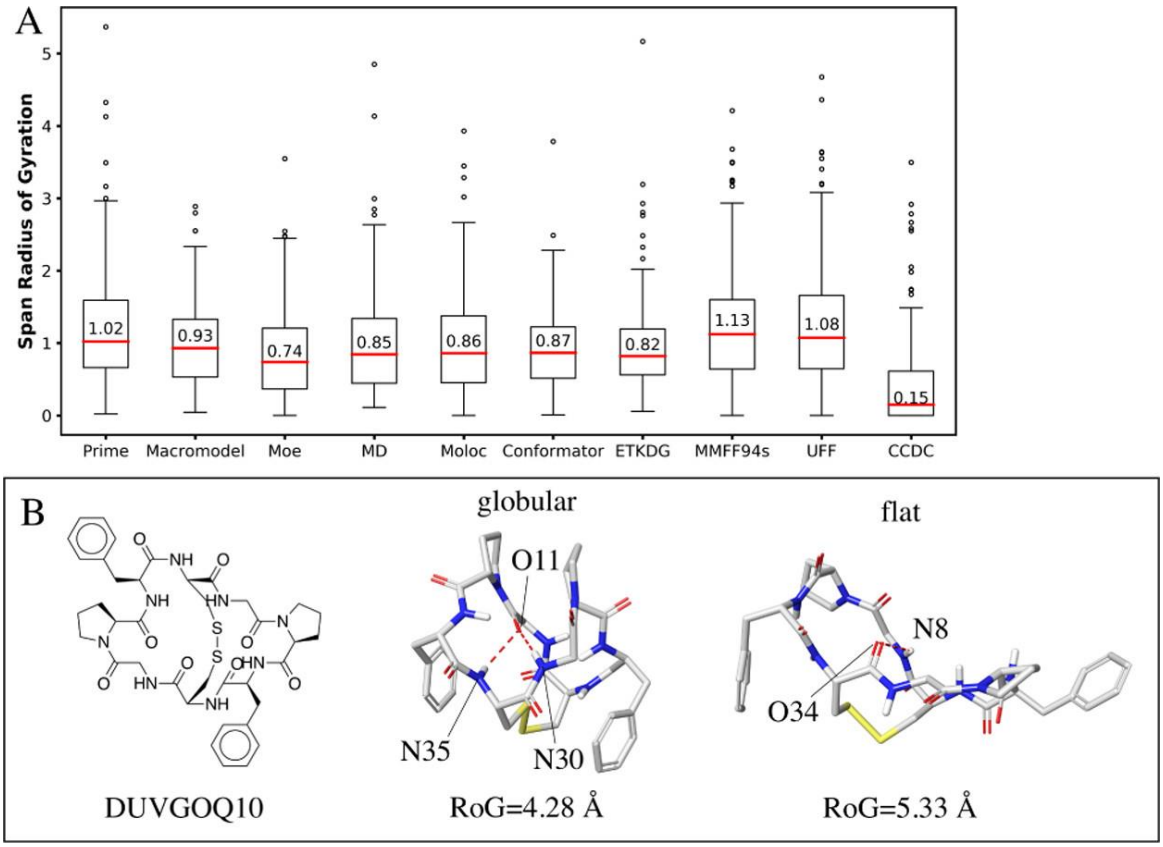


Figure 4. (A) Box plot of span RoG for each method and (B) example of a cyclic octapeptide61 in its globular (lowest RoG) and flat-like conformations (highest RoG) with intramolecular hydrogen bonds predicted with Moloc (red dotted lines).

Speed

Surprisingly, the speed of macrocyclic conformation generation differed dramatically between the software ranging from seconds to more than a day. This will have consequences for usage in virtual screening of large macrocycle libraries. Because sampling is carried out under similar conditions, comparisons allow analysis of the time required to accomplish the conformational task. The overall results of the computational speed are shown in Figure 5. With 2.6 s per entry, CCDC conformer generator outperformed the other software in time needed to finish a conformational job. On the other hand, MD was the slowest followed by conformator, which required 17.9 h. Prime, Moloc, and MOE produced conformations with a similar speed within 1 h with nonsignificant differences between MOE and Moloc (Table S5). More interestingly, we observed a statistical difference between ETKDG alone and UFF/MMFF94s resulting in a median of 35.1 s, 1.3 min, and 17.6 per entry.

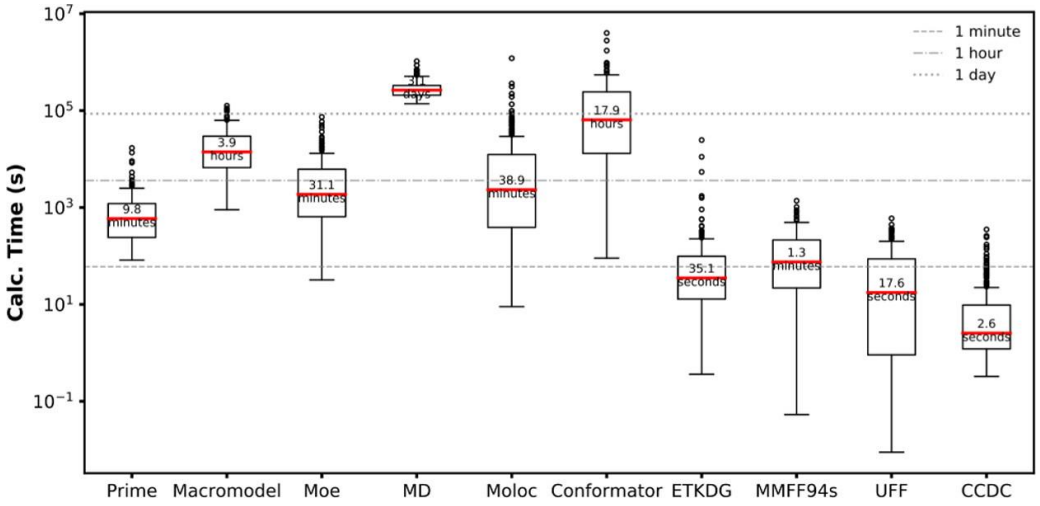


Figure 5. Box plot showing the distribution of the speed ranges for each entry. The reader is referred to Figure 2 for the legend. Three significant threshold values were added to visualize the differences in the performance level in completing a conformation work, i.e., 1 min, 1 h, and 1 d.

Study Cases

In addition to the benchmark results described above, we report cases of effective accuracy in predicting the crystallographic coordinates of macrocycles using Moloc both in terms of lowest RMSDbackbone/ RMSDheavy atoms and in relation with the ring size. For convenience, we kept the same categories as previously reported [49], binning the database in three groups containing 10−19, 20−29, and over 30 ring atoms, respectively. We referred to Prime as a comparative example among other commercial software.

10−19-Ring-Sized Macrocycles

10−19-ring-sized macrocycles represent a challenge in the context of organic synthesis because of the high energetic strain. Similarly, medium-sized rings suffer from increased ring strain over their 5- and 6- membered or macrocyclic congeners [62,63]. This can be quantitatively captured in deviations from ideal antiperiplanar conformations, transannular strain, and Pitzer strain components. Out of the total 208, 117 macrocycles belong to this class, including 30 from PDB, 79 from CSD, and 8 from BIRD datasets. According to our findings, Moloc predicted the coordinates of ACOPUF (Figure 6A), a 12-ring-sized macro- cycle from the CSD database, with an RMSDbackbone of 0.07 Å -slightly better than Prime (0.12 Å)- and with less conformations (requiring only 93 for the former against 871 for the latter). In a similar fashion, Moloc predicted the bioactive conformation of cytochalasin D (Figure 6C), an 11- membered ring macrocycle from the PDB database, with a high accuracy (0.12 Å) employing only 9 conformers, whereas Prime

(0.15 Å) employed 185. BANROX (Figure 6B) and DOZWUL (Figure 6D) were two CSD macrocycles of 13- and 14-atom backbone, respectively, with an RMSDheavy atoms of 0.09 and 0.10 Å. These data indicate that this software is highly accurate for medium-sized rings. In contrast to Prime, Moloc also proved to be superior in terms of the number of conformations, producing only 33 and 93 conformers rather than 95 for BANROX and 388 for DOZWUL, and accuracy with RMSDheavy atoms values of 0.44 and 0.41 Å for Prime.

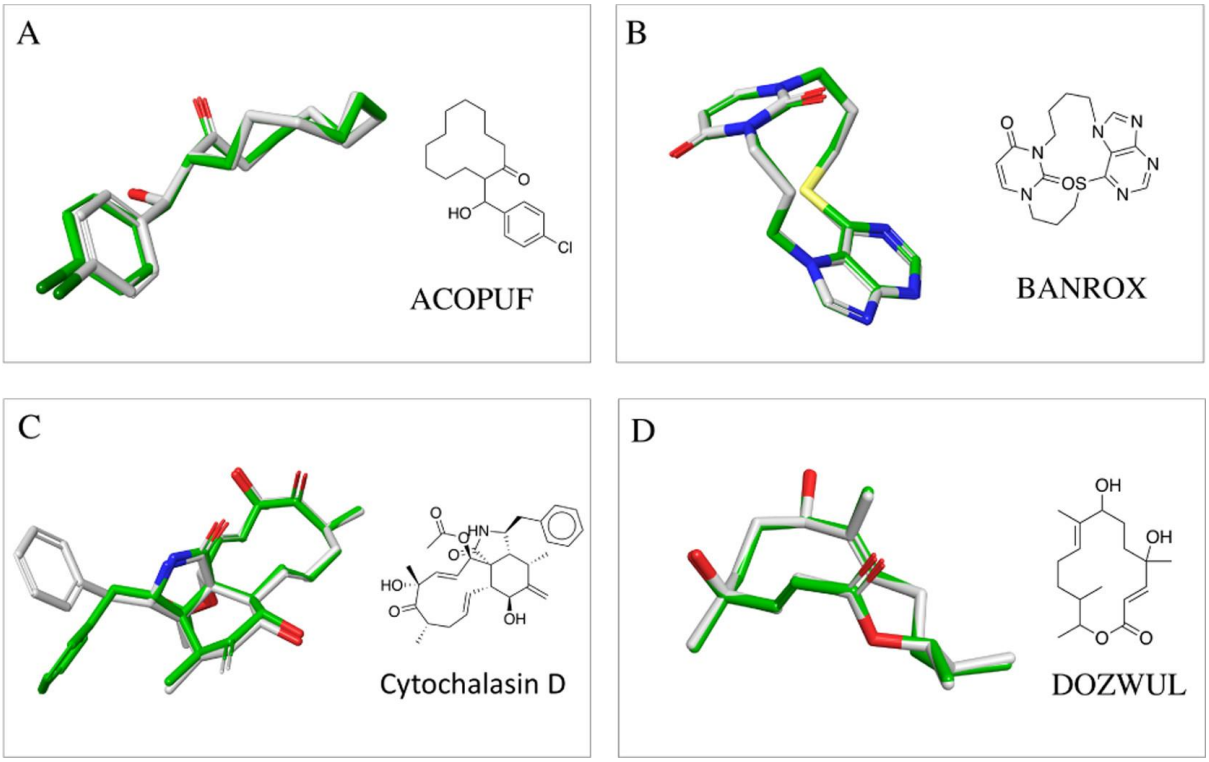


Figure 6. Examples of macrocycles having a flexibility of 10−19-atom backbone and indication by their dataset identifier (A−D). The atoms of the crystallographic structure to which the lower RMSD conformer has been aligned are colored in gray, whereas those of the conformer predicted using Moloc are in green.

20−29-Ring-Sized Macrocycles

This category includes 67 X-ray structures, 27 from PDB, 34 from CSD, and 6 from BIRD database. On the one hand, Moloc reproduced 7 entries with high accuracy (<0.5 Å) and 38 with accuracy <1.0 Å, with the best being DEMJAG10 (Figure 7A) and kabiramide C (Figure 7B), two macrocycles of 22 and 25 ring size from the CSD and PDB dataset, whose closest coordinates to the bioactive molecule were 0.13 and 0.17 Å RMSDbackbone, respectively. Despite producing 789 and 172 conformations, Moloc remained superior to Prime, for which the closest coordinates for the two referred macrocycles were 0.82 and 0.35 Å, respectively (1000 conformations per entry). On the other hand, it is also interesting to assess the robustness of Moloc in generating

accurate conformations of the heavy atoms. In that respect, only 11 crystal structures resulted in an interval of RMSDheavy atoms < 1.0 Å mostly belonging to the CSD (10) with only one from the PDB dataset (Figure 7C). Among these macrocycles, it is noteworthy to mention WURVEL (Figure 7D), a 27-membered ring entry from the CSD database, whose closest atomic coordinates (1.0 Å) indeed were not dissimilar from those predicted using Prime (1.06 Å); nevertheless, Moloc produced 163 conformations while Prime produced 983.

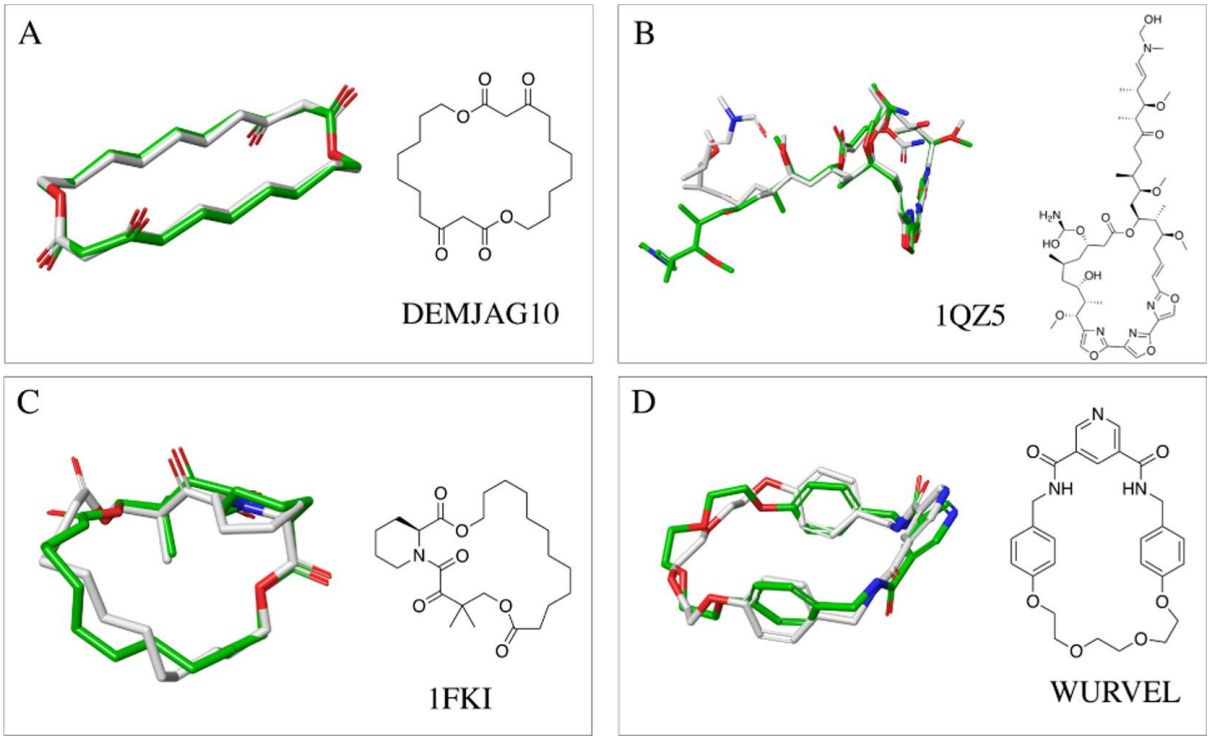


Figure 7. Examples of macrocycles having a flexibility of 20−29-atom backbone and their dataset identifier (A−D). The atoms of the crystallographic structure to which the lower RMSD conformer has been aligned are colored in gray, whereas those of the conformer predicted using Moloc are in green.

>30-Ring-Sized Macrocycles

Highly flexible macrocycles represent a challenge for every conformational algorithm, given the large number of rotatable bonds and possible values of torsional angles around the ring. Another problem is the number of replacements that attach to the ring and their degree of branching. In this subset, a total of 24 crystalline structures can be found and, specifically, 5 are cross-linked and another 5 are cyclopeptides that were originally included by the Prime developers in order to make the benchmark more challenging. Five macrocycles, all belonging to the CSD database, appeared in the list predicted with RMSDbackbone < 1.0 Å. Among them, Moloc predicted the crystallographic coordinates of OCERET (Figure 8A), a 35-atom backbone macrocycle, with an

RMSDbackbone of 1.04 Å with 168 conformations. On comparison, Prime performed slightly better with 0.83 Å but produced 957 conformations. Only SUMMOC (Figure 8B) and LENPEA (Figure 8C) were predicted below the threshold of 1.0 Å with values of RMSDheavy atoms of 0.74 and 0.92 Å, respectively. In addition to the advantage of Moloc being able to handle large-sized macrocycles, we noticed a limitation of Moloc in the complexity of the functional groups expressed in terms of degree of branching. An example of this limit is shown in Figure 8D. The measured RMSDheavy atoms of (−)-rhizopodin (PDB: 2VYP), a potent actin-binding anticancer molecule [64], decreases from 6.444 to 1.49 Å upon pruning the lateral substituents. This evidence can be explained by the ability of Prime to randomly cleave the macrocycle and reconnect the two generated semiloops.

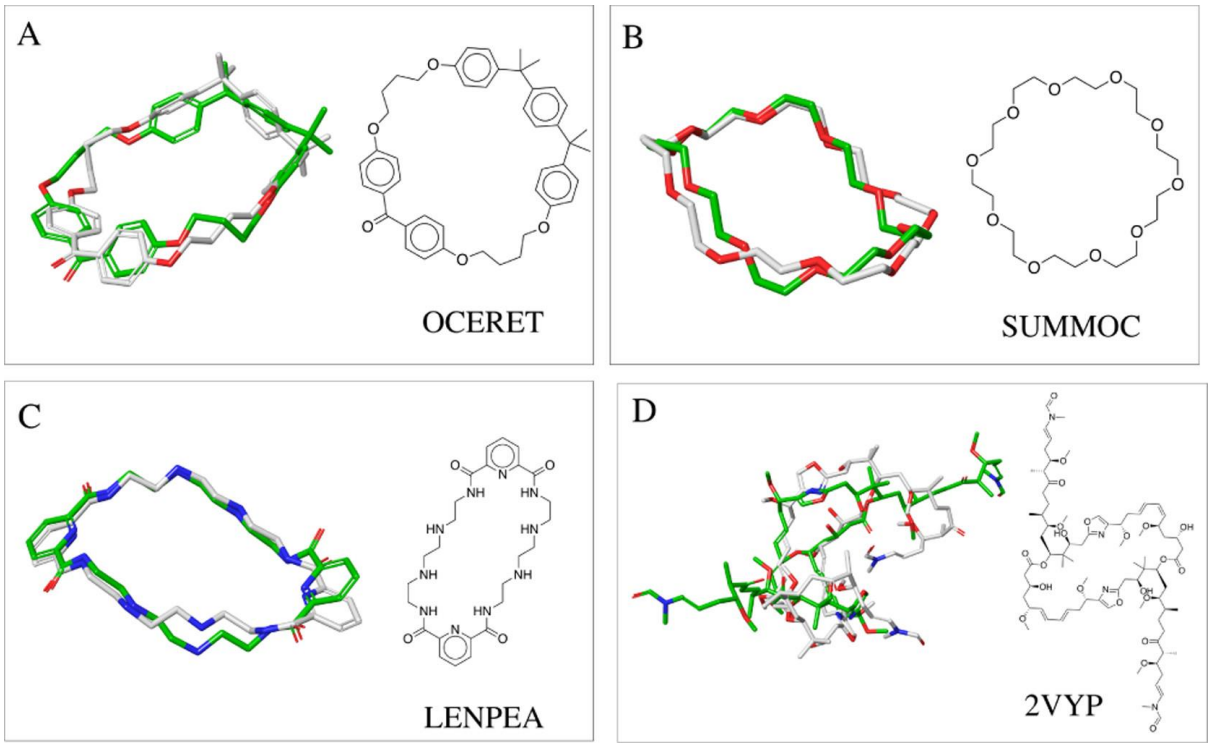


Figure 8. Examples of macrocycles indicated by their dataset identifier (A–D). The atoms of the crystallographic structure to which the lower RMSD conformer has been aligned are colored in gray, whereas those of the conformer predicted using Moloc are in green.

Intramolecular Interactions

The ideal software is required to predict intramolecular interactions as it is generally appreciated that they play a pivotal role in defining both overall shape of a molecule [65] and the stabilization of the functional groups by masking or exposing them to the external environment [66]. This change regulates the passive membrane permeability of macrocycles which adopt a globular

shape while passing through the lipidic environment of the membrane and adopt a stretched conformation in the cytosol/ extracellular environment [45]. Knowledge of the chameleonic properties of macrocycles has recently expanded far beyond the historical case of ciclosporin A [67,68]. As exemplified by the crystal structures of cyclosporin A in chloroform (CSD ID $P2_12_12_1$) and in the protein bound form (PDB ID: 2X2C [69]), the conformational change is followed by the formation of new intramolecular hydrogen bonds, underlying their role in the dynamics of binding. As can be seen in Figure 9A, the crystal structure of CUQYUI, the 24- atoms backbone of the non-cross-linked cyclopeptide has 4 internal hydrogen bonds (between N15 and O2, N16 and O2, and O6 and N11 as well as one transannular interaction between N12 and O10). Moloc successfully predicted three of these internal hydrogen bonds with an RMSDheavy atoms of 1.365 Å and, most notably, matched the lowest global minimum among the 38 local minima, with a potential energy of 5.33 kcal/mol. 3WNF-ACE (Figure 9B) is a 20-atom backbone hexacyclic peptide whose binding affinity for HIV-1 integrase was measured in the low millimolar range by surface plasmon resonance and HSQC-NMR while the binding mode with the target was confirmed by X-ray crystallography [70]. Visual inspection of the cocrystal structure revealed the presence of two internal hydrogen bonds between N35 and O13, and N10 and O38 and two transannular interactions, between O34 and N27, and O2 and N10. Moloc was able to predict three of these four interactions with reasonable accuracy (RMSDheavy atom = 1.945 Å) and a local minimum with a potential energy of 11.13 kcal/mol. YIWHOB01 (Figure 9C) is a 30-atom backbone non-cross-linked artificial macrocycle used as a charge transfer system in the field of supramolecular chemistry [71]. Visual inspection of the CSD structure revealed the presence of a π-stacking interaction between the pyridine and phenyl rings. Again, Moloc predicted the conformation with the bipyridinium units being parallel to the phenyl ring with an RMSDheavy atom of 1.642 Å and a potential energy of 9.846 kcal/mol, despite minor deviations at the dioxoaryl moiety.

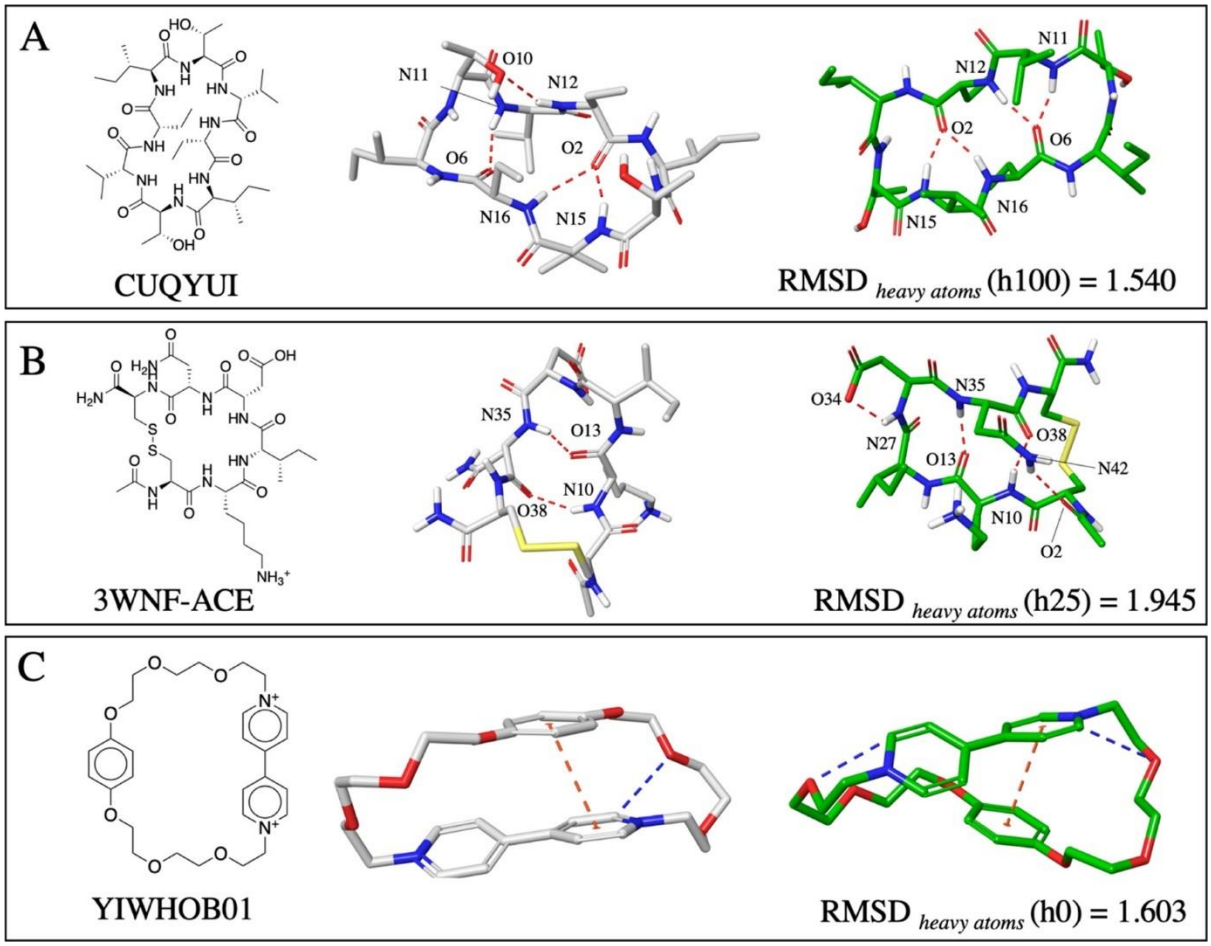


Figure 9. Panel showing the intramolecular interactions predicted using Moloc (green sticks) for (A) CUQYUI, (B) 3WNF-ACE, and (C) YIWHOB0 alongside with the RMSDheavy atoms calculated for the hydrogen bond weight applied in the MAB force field. Hydrogen bonds, π- stacking, and aromatic hydrogen bonds are, respectively, colored as red, blue, and orange dotted lines while the crystal structure atoms are represented as gray sticks.

User-Defined Energy Threshold for Improved Accuracy and Diversity

In a standard Moloc conformational job, the structures are only kept if their energy is less than 10 kcal/mol above the lowest-energy conformation. Such an energetic cutoff is typical for many other conformational software. However, Prime sets the cutoff to 100 kcal/mol. Thus, we have quantified the diversity and the accuracy at 100 kcal/mol and chose 4MNW and 4KEL, two cyclopeptides, cross-linked macrocycles with 42-atom backbone. Based on our data (Table S6), no improvement over the diversity was observed independently from the chosen threshold because the number of unique fingerprints for 4MNW (192) and 4KEL (290) remained unchanged. However, when the energy threshold was increased to 100 kcal/mol, Moloc produced new conformers with expanded globularity because the span RoG increased from 1.179 to 1.660 Å for 4KEL and from 1.041 to 1.704 Å for 4MNW. Additionally, we observed a marginal improvement in both the ring and the heavy atom structure accuracies: −0.42 Å/−0.23

Å (4MNW) and −0.22 Å/−0.08 Å (4KEL) at 20 kcal/mol and −0.83 Å/−0.76 Å (4MNW) and −0.25 Å/−0.39 Å (4KEL) at 100 kcal/mol (Figure S2A). As the number of conformations for both cases exponentially increased (Figure S2B), the global minimum energy of the most accurate conformer of 4MNW displays an increase in the potential energy by 6 and 15 kcal/mol, whereas for 4KEL, the equivalent values were 8 and 5 kcal/mol (Figure S2C, D).

Discussion

Computational screening of large virtual macrocycle libraries is an effective way to prioritize compounds for expensive and time-consuming synthesis in the laboratory. We have recently described convergent and short syntheses of macrocycles using MCR. One synthesis consisted of a short two-step assembly of macrocycles from cyclic anhydrides, diamines, oxo components (aldehydes and ketones), and isocyanides. Based on commercial availability of the building blocks, a very large chemical space is spanned: 20 (cyclic anhydrides) × 20 (diamines) × 1000 oxo components × 1000 isocyanides = 400 million macrocycles. Computational generation of conformers for such a large chemical space requires fast and optimized software. Therefore, in this manuscript, we have benchmarked Moloc versus available commercial and freeware for their performance as defined by accuracy, speed, exhaustiveness, diversity, and SE. Our results confirmed that Prime, MM, and MOE possess higher accuracy in reproducing both the heavy atoms and ring coordinates of the crystallographic macrocycle references. According to our results, conformational sampling with ETKDG algorithm could be improved by subsequent minimizations steps with MMFF94s but not UFF. This finding could be related to the existence of out-of-plane bending and dihedral torsion parameters to planarize certain types of delocalized trigonal N atoms applied by the MMFF94s force field, thus providing a better match to the reference crystal structures. However, UFF contains basic parameters for all types of atoms on hybridization and connectivity and thereby is able to parameterize the restricted patterns of dihedral angles and rotatable bonds, both present in macrocycles.44 Nevertheless, these data lead us to suggest that the implementation of minimization steps employing specific force fields after conformational sampling of macrocycles would lead to improvements of sampling. For instance, the OPLS-2005 in Prime or MAB force field in Moloc represent the most accurate commercial and open software, respectively. Such an evidence could allow further analysis to study the effect of different force fields to improve macrocycle sampling. On the other hand, we show that the use of DG methods as ETKDG could be improved to generate conformers closely related to the crystal structures. In this sense, a modification to the ETKDG algorithm for

macrocycle sampling has been recently published by the developer team of RDKIT and will be available in the upcoming RDKIT release 2020.03.47 Along with a restriction in search space for macrocycles, the new implementations in ETKDG will include additional torsional-angle potentials to describe small aliphatic rings and adapt the previously developed potentials for acyclic bonds to facilitate the sampling of macrocycles. Nevertheless, because of the novelty of this algorithm, more testing is needed to evaluate its capability in diverse and challenging macrocycle datasets, such as those presented in this work. MD was performed only under solvated conditions49 with no major improvement in generating high-quality conformers according to the SE value. However, other reported MD-based approaches using different simulation conditions have reported the importance of solvation for the generation of bioactive conformations of macrocycles [72]. An enhanced sampling method has been reported using MD simulations that resulted in a reliable method to reproduce the experimentally determined structure of three macrocycles [73]. Nevertheless, the major drawback for MD-based methods relies on its low scalability of large and diverse macrocycle datasets. As a result, such methods can be an option when working with a limited number of macrocyclic structures but not for virtual screening approaches such as Prime, MM, Moloc, ETKDG, or other software reported here. Although CCDC conformer generator was one of the most efficient software for conformer generation in terms of speed and exhaustiveness, it suffers a low rate of conformational sampling exploration as only one single conformer was generated for 37 structures. The most noticeable exception relies on 76 cases where the RMSDbackbone values were unrealistically lower than 0.1 Å and hence equal to the crystallographic reference. This behavior could be explained by a bias in the sampling of entries from CSD: the CCDC conformer generator assigns the crystallography coordinates prior to conformation sampling. The CCDC conformer generator uses bond lengths and valence angles taken from CCDC Mogul and one of its best strengths consists in the use of dynamic rotamer libraries that are automatically updated with new data inside of CCDC [74,75]. However, although CCDC conformer generator has implemented strategies to deal with conformer generation of rings as set preclustered templates for isolated, fused, spiro-linked, and bridged ring systems [75], there is no specific method regarding macrocyclic conformers yet described. For instance, in rings for which no template is obtainable from Mogul data, the templates are generated on the fly using rotamer distributions for cyclic bonds [74,75]. If ring generation fails and no template structure can be generated, the ring conformation from the 3D input structure is used. According to our results, the conformational sampling with CCDC conformer generator for the CSD entries, bond lengths, and valence angles were taken from CCDC Mogul retrieving conformers with conformations close to the crystal structures. Thus, for the macrocycles not included in CSD database, the conformers were generated either from an on-the-fly template assignment or using

the input coordinates. This could explain the lowest number of conformers generated per entry and the reduced number of unique TFs. Furthermore, the span in RoG values from CCDC conformer generator suggests a tendency to retain conformations with higher compaction in comparison with any other methods for macrocycle conformational sampling described here, thus omitting possible extended states. Taking these results together, the restricted usage of CCDC conformer generator within the macrocycle conformational sampling could lead to poor results in terms of conformational space exploration or even a lack of conformers, suggesting that this tool is useful only to generate conformers for small molecules or for the assignment of crystallographic coordinates to macrocycle structures. Overall, our analysis indicated conformator as the lowest efficiency conformational sampling software tested in this work. This tool showed one of the lowest exhaustiveness values among the studied methods, just below that of MD. The accuracy of conformator reproducing the macrocycle backbone is also the lowest and is also one of the slowest conformational sampling methods generating structures with the lowest span in RoG of all methods tested. Nevertheless, the authors of conformator have tested this algorithm employing 49 different macrocyclic structures [46]. These evidence suggest that the use of conformator could be restricted to small-to-medium macrocycles. Further analysis and testing are needed to assess the feasibility of conformator in generating conformers for a dataset containing large and complex structures. Furthermore, this software produces conformations that differ from each other by rotation of one single bond at a time which may limit its use to macrocycle with few rotatable bonds. As for Moloc, we are indeed aware that reproducing the accuracy of all heavy atoms, as our RMSDheavy atoms data demonstrate, represents its main limitation. However, we would like to emphasize that one of the main challenges in the conformational analysis of macrocycles is the accuracy of ring atoms. Based on our RMSDbackbone data, Moloc has a similar accuracy to the negative control (MD) and MD, Moloc, and ETKDG alone or in combination with MMFF94s, implying that it can be used as a valid alternative to these two methodologies to produce conformations with a similar accuracy. Most importantly, Moloc retains good exhaustiveness, SE, and economy in terms of least numbers of conformers to generate high quality conformers without requiring 1000 or more conformers for the exhaustive exploration of the chemical space, saving computational resources and avoiding redundancy in the conformers generated, suggesting this software as an acceptable alternative to Prime, MM, and MD for sampling. One major drawback of Moloc is that it relies on the number of symmetry elements within the macrocycle structure needed for the sampling. This is particularly evident in the case of POGLIH, a macrocycle from the CSD, for which 5 days were necessary to complete the conformational sampling. Indeed, the enumeration of topological symmetries is intended to avoid the counting of identical conformations that vary only by altered atomnumbering (e.g., 180°

rotation of a phenyl ring in the structure). Such enumeration takes an (exponentially) increasing time in accordance with the number of symmetry elements. For POGLIH, all 8 phenyl rings can be rotated, and methyl groups can be exchanged, as well as oxygen in the sulfates. In addition, the whole structure has a twofold symmetry. All in all, there are over 32,000 symmetry elements present, meaning that the same conformation may occur 32,000 times indicating that a threshold or restricted search of symmetries and their calculation could improve the speed of sampling. Another limitation of Moloc consists in sampling macrocycles with complex side chains: this has been seen in rhizopodin (PDB: 2VYP), a potent actin-binding anticancer agent.64 Aiming to understand the relation between the accuracy and the side-chain complexity, we first trimmed the two 15-atom-branched symmetrical side chains of rhizopodin and subsequently sampled again the macrocycle (Figure S1). As a result, we observed an improvement of heavy atom accuracy (from 6.27 to 2.17 Å) and an increased number of conformers (increasing from 62 to 205). Nevertheless, several parameters allow the user a full control of the output ensembles, making Moloc a flexible piece of software for the molecular modeling of macrocycles. Our data indicate that the number of ensembles can be interactively controlled by applying either by energy thresholds (parameter “e”) or hydrogen bound weight (parameter “h”) term in the batch mode, allowing the enumeration of globular or flat conformations, the identification of intramolecular hydrogen bonds, and potentially predicting the most accurate ones in nonpolar environments. Taken altogether, these applications of Moloc indeed represent a “nice-to-have” tool in the molecular modeling toolkit of permeable macrocycles. Not lastly, the user can decide whether to apply a final energy minimization after conformational sampling followed by the addition of hydrogens to heteroatoms by invoking the parameter “q1”. As a result, Moloc returns all the energetic components calculated by MAB per conformer produced, bonds, valence angles, torsions, pyramidalities, 1−4 repulsion, van der Waals interactions, hydrogen bonds, and polar repulsion. To our knowledge, recent algorithms were published with already built-in protocols including the maximum ensemble size, RMSD or energy thresholds, and further constrains such as NMR data, enforcement of the chirality, geometry check before sampling, and application of a filter to retain the conformers according to a certain R value of the crystal structures [38,46,49,76]. MM presents indeed the advantage of tuning several parameters such as electrostatic treatment and possibility to choose two different force fields (OPLS-2005 or MMFF94s) [39]. In the case of open-access software, such as ETKDG, recently, new improvements were released in order to favor certain interactions or orientation angles [48]. Additionally, we would like to point out that CCDC conformer generator as well as ETKDG and conformator are knowledge-based systems with pre-existing rotational libraries of small-medium rings. This implies that if a test set entry is derived from the CSD, it will have prior information and

make use of these coordinates. Nevertheless, CSD entries were retained in knowledge-based systems. Finally, a possible strategy to improve the accuracy of complex macrocycles could be the implementation of further shape constrains accounting for the crystallographic packing forces because most of the macrocyclic crystal structures are flattened in a high-energy level conformation. Additional improvement of Moloc should also consider the flexibility of the complex side chains because the current version of the algorithm starts the identification of the first generic shape from a polar coordinate of a circle with an acceptable degree of accuracy and time.

Conclusions

In this work, we have benchmarked the shape-guided algorithm using a dataset of 208 macrocycles from Prime publication, carefully selected on the basis of structural complexity (e.g., ring size, cyclopeptide/aliphatic, cross-linkings) and we have quantified accuracy, diversity, speed, exhaustiveness, and SE with four conformational commercial (Prime, MM, MOE, and MD) and five open-access (ETKDG, MMFF94s, UFF, CCDC, and conformator) software packages. A Python script to streamline the whole data collection of these parameters has been written ad hoc. The results of our benchmark are summarized in Table 4. Although Prime, MM, MOE, and MD remained the most accurate software tested in this paper in reproducing macrocycle heavy atoms, Moloc retained the same exhaustiveness. However, Moloc stood out for the highest SE in producing an acceptable number of conformations per entry and three-quarters of the database were processed with high accuracy (RMSDbackbone < 1.0 Å). Interactive control of the hydrogen bond terms allows the enumeration of globular and flat conformers and prediction of intramolecular interaction in a nonpolar solvent. However, the structural accuracy of Moloc is hampered by long-branched side chains. In that respect, side chain pruning in the batch mode with “Mdfy”, a built-in module within Moloc, and subsequent reattachment to the ring could be an option for future improvement. Surprisingly, minimization with UFF and MMFF94s managed to produce macrocycles with the most diverse shapes in terms of RoG, suggesting these types of software as a valid free alternative for the prediction of the most likely shape that the macrocycles can adopt in their bulk environment, for example, the cellular membrane or water. Follow-up studies could include modifications to ETKDG algorithm or the use of force field minimization in order to predict the X-ray structure. For instance, the evaluation of ETDKG conformational sampling was combined with OPLS- 2005 and/or MAB as minimization methods.

Table 4. Summary Table of the Benchmark

Method	Prime	MM	MOE	MD	Moloc	conformator	ETKDG	MMFF94s	UFF	CCDC
$RMSD_{heavy\ atoms}$ (Å)	0.878	0.655	0.765	1.052	1.910	1.990	2.165	1.793	2.083	2.067
$RMSD_{backbone}$ (Å)	0.396	0.383	0.417	0.562	0.652	0.801	0.743	0.668	0.766	0.476
N of conformations	972	300	76	1000	67	338	1000	998	535	8
TF	707	100	48	59	67	338	1000	998	535	8
span RoG (Å)	1.02	0.93	0.74	0.85	0.86	0.87	0.82	1.13	1.08	0.15
exhaustiveness	1.00	1.00	1.00	1.00	1.00	0.91	1.00	0.95	0.95	1.00
SE	0.76	0.33	0.63	0.06	1.00	0.66	1.00	0.95	0.95	0.75
speed	9.8 min	3.9 h	31.1 min	3.1 d	38.9 min	17.9 h	35.1 s	1.3 min	17.6 s	2.6 s

Supplementary information

Table S1 summary of the pairwise Krustal-Wallis H-test calculated for the median of RMSD $_{\text{heavy atoms}}$ computational sampling methods reported. * $p \leq 0.05$, ** $p \leq 0.01$, *** $p \leq 0.001$, ns: not significant.

Comparison	p-value	statistical significance
Conformator_vs_CCDC	0,1231	ns
Conformator_vs_ETKDG	0,4009	ns
Conformator_vs_MMFF94s	0,5512	ns
Conformator_vs_UFF	0,344	ns
ETKDG_vs_CCDC	0,0507	ns
ETKDG_vs_MMFF94s	0,1264	ns
ETKDG_vs_UFF	0,967	ns
MD_vs_CCDC	0,0011	**
MD_vs_Conformator	<0,001	***
MD_vs_ETKDG	<0,001	***
MD_vs_MMFF94s	<0,001	***
MD_vs_Moloc	<0,001	***
MD_vs_UFF	<0,001	***
MMFF94s_vs_CCDC	0,2774	ns
MMFF94s_vs_UFF	0,1002	ns
MOE_vs_CCDC	<0,001	***
MOE_vs_Conformator	<0,001	***
MOE_vs_ETKDG	<0,001	***
MOE_vs_MD	0,0057	**
MOE_vs_MMFF94s	<0,001	***
MOE_vs_Moloc	<0,001	***
MOE_vs_UFF	<0,001	***
Macromodel_vs_CCDC	<0,001	***
Macromodel_vs_Conformator	<0,001	***
Macromodel_vs_ETKDG	<0,001	***
Macromodel_vs_MD	<0,001	***
Macromodel_vs_MMFF94s	<0,001	***
Macromodel_vs_MOE	0,9174	ns
Macromodel_vs_Moloc	<0,001	***
Macromodel_vs_UFF	<0,001	***
Moloc_vs_CCDC	0,3281	ns
Moloc_vs_Conformator	0,3895	ns
Moloc_vs_ETKDG	0,111	ns
Moloc_vs_MMFF94s	0,833	ns
Moloc_vs_UFF	0,1025	ns
Prime_vs_CCDC	<0,001	***

Prime_vs_Conformator	<0,001	***
Prime_vs_ETKDG	<0,001	***
Prime_vs_MD	0,0091	**
Prime_vs_MMFF94s	<0,001	***
Prime_vs_MOE	0,738	ns
Prime_vs_Macromodel	0,2048	ns
Prime_vs_Moloc	<0,001	***
Prime_vs_UFF	<0,001	***
UFF_vs_CCDC	0,0474	*

Table S2 summary of the pairwise Krustal-Wallis H-test calculated for the median of $RMSD_{backbone}$ computational sampling methods reported. * $p \leq 0.05$, ** $p \leq 0.01$, *** $p \leq 0.001$, ns: not significant.

Comparison	p-value	statistical significance
Conformator_vs_CCDC	<0.001	***
Conformator_vs_ETKDG	0.6258	ns
Conformator_vs_MMFF94s	0.0102	*
Conformator_vs_UFF	0.6885	ns
ETKDG_vs_CCDC	<0.001	***
ETKDG_vs_MMFF94s	0.0269	*
ETKDG_vs_UFF	0.8099	ns
MD_vs_CCDC	0.0287	*
MD_vs_Conformator	<0.001	***
MD_vs_ETKDG	<0.001	***
MD_vs_MMFF94s	0.0103	*
MD_vs_Moloc	0.0615	ns
MD_vs_UFF	<0.001	***
MMFF94s_vs_CCDC	0.0023	**
MMFF94s_vs_UFF	0.0136	*
MOE_vs_CCDC	0.3210	ns
MOE_vs_Conformator	<0.001	***
MOE_vs_ETKDG	<0.001	***
MOE_vs_MD	<0.001	***
MOE_vs_MMFF94s	<0.001	***
MOE_vs_Moloc	<0.001	***
MOE_vs_UFF	<0.001	***
Macromodel_vs_CCDC	0.7173	ns
Macromodel_vs_Conformator	<0.001	***
Macromodel_vs_ETKDG	<0.001	***
Macromodel_vs_MD	<0.001	***
Macromodel_vs_MMFF94s	<0.001	***
Macromodel_vs_MOE	0.7203	ns

Macromodel_vs_Moloc	<0.001	***
Macromodel_vs_UFF	<0.001	***
Moloc_vs_CCDC	0.0034	**
Moloc_vs_Conformator	0.0018	**
Moloc_vs_ETKDG	0.0036	**
Moloc_vs_MMFF94s	0.4101	ns
Moloc_vs_UFF	0.0016	**
Prime_vs_CCDC	0.5943	ns
Prime_vs_Conformator	<0.001	***
Prime_vs_ETKDG	<0.001	***
Prime_vs_MD	<0.001	***
Prime_vs_MMFF94s	<0.001	***
Prime_vs_MOE	0.9361	ns

Table S3 summary of the pairwise Krustal-Wallis H-test calculated for the torsional fingerprint median of the computational sampling methods reported. * $p \leq 0.05$, ** $p \leq 0.01$, *** $p \leq 0.001$, ns: not significant.

Comparison	p-value	statistical significance
Prime_vs_Macromodel	<0.001	***
Prime_vs_Moe	<0.001	***
Prime_vs_MD	<0.001	***
Prime_vs_Moloc	<0.001	***
Prime_vs_Conformator	<0.001	***
Prime_vs_ETKDG	<0.001	***
Prime_vs_MMFF94s	<0.001	***
Prime_vs_UFF	0.4048	ns
Prime_vs_CCDC	<0.001	***
Macromodel_vs_Moe	<0.001	***
Macromodel_vs_MD	<0.001	***
Macromodel_vs_Moloc	<0.001	***
Macromodel_vs_Conformator	<0.001	***
Macromodel_vs_ETKDG	<0.001	***
Macromodel_vs_MMFF94s	<0.001	***
Macromodel_vs_UFF	<0.001	***
Macromodel_vs_CCDC	<0.001	***
Moe_vs_MD	0.6715	ns
Moe_vs_Moloc	0.1801	ns
Moe_vs_Conformator	<0.001	***
Moe_vs_ETKDG	<0.001	***
Moe_vs_MMFF94s	<0.001	***
Moe_vs_UFF	<0.001	***

Moe_vs_CCDC	<0.001	***
MD_vs_Moloc	0.5448	ns
MD_vs_Conformator	<0.001	***
MD_vs_ETKDG	<0.001	***
MD_vs_MMFF94s	<0.001	***
MD_vs_UFF	<0.001	***
MD_vs_CCDC	<0.001	***
Moloc_vs_Conformator	<0.001	***
Moloc_vs_ETKDG	<0.001	***
Moloc_vs_MMFF94s	<0.001	***
Moloc_vs_UFF	<0.001	***
Moloc_vs_CCDC	<0.001	***
Conformator_vs_ETKDG	<0.001	***
Conformator_vs_MMFF94s	<0.001	***
Conformator_vs_UFF	0.0029	**
Conformator_vs_CCDC	<0.001	***
ETKDG_vs_MMFF94s	<0.001	***
ETKDG_vs_UFF	<0.001	***
ETKDG_vs_CCDC	<0.001	***
MMFF94s_vs_UFF	<0.001	***
MMFF94s_vs_CCDC	<0.001	***
UFF_vs_CCDC	<0.001	***

Table S4 summary of the pairwise Krustal-Wallis H-test calculated for the medians' span radius of gyration across the computational sampling methods reported. * $p \leq 0.05$, ** $p \leq 0.01$, *** $p \leq 0.001$, ns: not significant.

Comparison	p-value	statistical significance
Prime_vs_Macromodel	0.0334	*
Prime_vs_Moe	<0.001	***
Prime_vs_MD	0.0014	**
Prime_vs_Moloc	0.0040	**
Prime_vs_Conformator	0.0056	**
Prime_vs_ETKDG	0.0016	**
Prime_vs_MMFF94s	0.5699	ns
Prime_vs_UFF	0.7871	ns
Prime_vs_CCDC	<0.001	***
Macromodel_vs_Moe	0.0050	**
Macromodel_vs_MD	0.2322	ns
Macromodel_vs_Moloc	0.3995	ns
Macromodel_vs_Conformator	0.4621	ns
Macromodel_vs_ETKDG	0.3470	ns
Macromodel_vs_MMFF94s	0.0071	**

Macromodel_vs_UFF	0.0201	*
Macromodel_vs_CCDC	<0.001	***
Moe_vs_MD	0.0837	ns
Moe_vs_Moloc	0.0805	ns
Moe_vs_Conformator	0.0258	*
Moe_vs_ETKDG	0.0171	*
Moe_vs_MMFF94s	<0.001	***
Moe_vs_UFF	<0.001	***
Moe_vs_CCDC	<0.001	***
MD_vs_Moloc	0.8531	ns
MD_vs_Conformator	0.5334	ns
MD_vs_ETKDG	0.5983	ns
MD_vs_MMFF94s	<0.001	***
MD_vs_UFF	0.0013	**
MD_vs_CCDC	<0.001	***
Moloc_vs_Conformator	0.8084	ns
Moloc_vs_ETKDG	0.9065	ns
Moloc_vs_MMFF94s	0.0011	**
Moloc_vs_UFF	0.0036	**
Moloc_vs_CCDC	<0.001	***
Conformator_vs_ETKDG	0.8560	ns
Conformator_vs_MMFF94s	<0.001	***
Conformator_vs_UFF	0.0027	**
Conformator_vs_CCDC	<0.001	***
ETKDG_vs_MMFF94s	<0.001	***
ETKDG_vs_UFF	<0.001	***
ETKDG_vs_CCDC	<0.001	***
MMFF94s_vs_UFF	0.7612	ns
MMFF94s_vs_CCDC	<0.001	***
UFF_vs_CCDC	<0.001	***

Table S5 summary of the pairwise Krustal-Wallis H-test calculated for the medians'speed across the computational sampling methods reported. * $p \leq 0.05$, ** $p \leq 0.01$, *** $p \leq 0.001$, ns: not significant.

Comparison	p-value	statistical significance
Prime_vs_Macromodel	≤0.001	***
Prime_vs_Moe	≤0.001	***
Prime_vs_MD	≤0.001	***
Prime_vs_Moloc	≤0.001	***
Prime_vs_Conformator	≤0.001	***
Prime_vs_ETKDG	≤0.001	***
Prime_vs_MMFF94s	≤0.001	***

Prime_vs_UFF	≤0.001	***
Prime_vs_CCDC	≤0.001	***
Macromodel_vs_Moe	≤0.001	***
Macromodel_vs_MD	≤0.001	***
Macromodel_vs_Moloc	≤0.001	***
Macromodel_vs_Conformator	≤0.001	***
Macromodel_vs_ETKDG	≤0.001	***
Macromodel_vs_MMFF94s	≤0.001	***
Macromodel_vs_UFF	≤0.001	***
Macromodel_vs_CCDC	≤0.001	***
Moe_vs_MD	≤0.001	***
Moe_vs_Moloc	0.5522	ns
Moe_vs_Conformator	≤0.001	***
Moe_vs_ETKDG	≤0.001	***
Moe_vs_MMFF94s	≤0.001	***
Moe_vs_UFF	≤0.001	***
Moe_vs_CCDC	≤0.001	***
MD_vs_Moloc	≤0.001	***
MD_vs_Conformator	≤0.001	***
MD_vs_ETKDG	≤0.001	***
MD_vs_MMFF94s	≤0.001	***
MD_vs_UFF	≤0.001	***
MD_vs_CCDC	≤0.001	***
Moloc_vs_Conformator	≤0.001	***
Moloc_vs_ETKDG	≤0.001	***
Moloc_vs_MMFF94s	≤0.001	***
Moloc_vs_UFF	≤0.001	***
Moloc_vs_CCDC	≤0.001	***
Conformator_vs_ETKDG	≤0.001	***
Conformator_vs_MMFF94s	≤0.001	***
Conformator_vs_UFF	≤0.001	***
Conformator_vs_CCDC	≤0.001	***
ETKDG_vs_MMFF94s	≤0.001	***
ETKDG_vs_UFF	≤0.001	***
ETKDG_vs_CCDC	≤0.001	***
MMFF94s_vs_UFF	≤0.001	***
MMFF94s_vs_CCDC	≤0.001	***
UFF_vs_CCDC	≤0.001	***

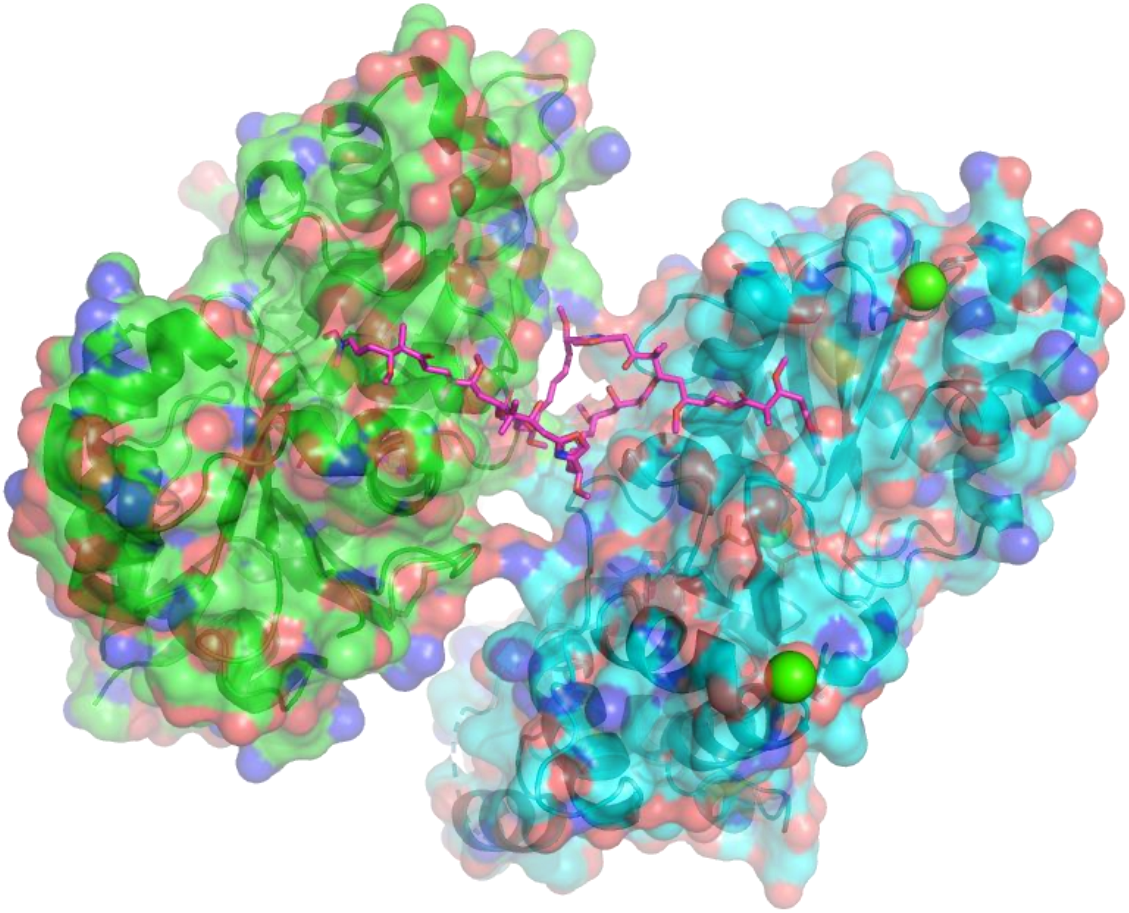


Fig.S1 Crystal structure of rhizopodin (magenta sticks) bound to two protein units of actin (green and cyan surface representation). Water and polyethylene glycol molecules were removed for clarity of visualization. Chloride atoms area represented as green spheres.

Table S6 Summary table of the parameters of Moloc at 100 kcal/mol energy threshold in comparison with commercial software. Nconf= number of conformations.

Entry	Method	N $_{conf}$	TF $_{backbone}$	TF	RoG (Å)	RMSD $_{heavy\ atoms}$ (Å)	RMSD $_{backbone}$ (Å)	MinEnergy
	Moloc	846	53	192	1.70	5.541	2.561	28.50
	Prime	7	7	7	1.50	5.107	2.045	74.64
4MNW_conf1	MM	207	98	98	1.05	5.118	2.475	0.00
	MOE	11	11	11	0.93	5.245	2.547	124.78
	MD	1000	528	528	1.64	4,646	2.263	17.35
	Moloc	802	52	200	1.66	3.740	2.037	45.36
	Prime	290	290	290	1.44	3.170	1.861	34.25
4KEL_conf1	MM	361	140	140	0.88	4.241	2.394	25.88
	MOE	4	3	3	0.25	4.649	2.685	39.29
	MD	1000	476	476	1.07	4.114	2.065	0.00

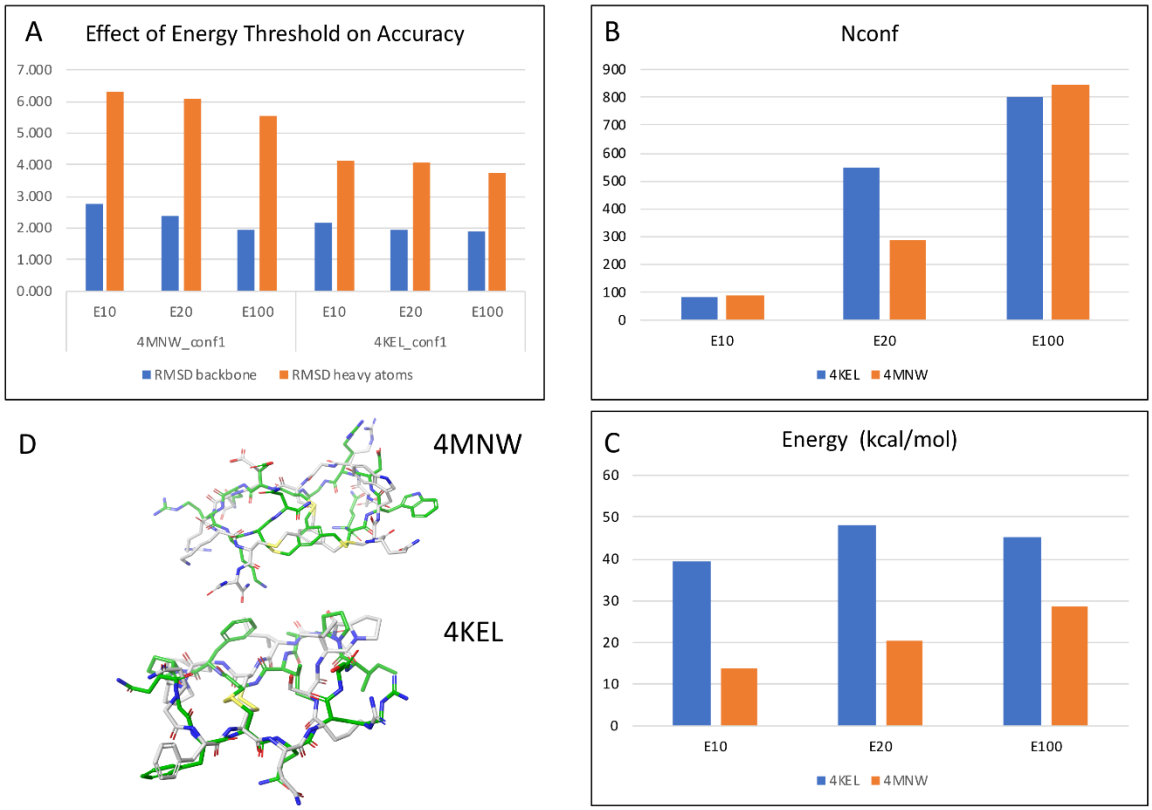


Fig.S2 Box plots showing the effects of different energy thresholds (10, 20 and 100 kcal/mol) over the (A) accuracy, (B) number of conformations and (C) local energy minimum. (D) Structural alignment between the lowest RMSD heavy atom conformer produced by Moloc (green stick) and the observed crystal structure (grey sticks) alongside with their PDB ID.

References

1. Frank AT, Farina NS, Sawwan N, Wauchope OR, Qi M, Brzostowska EM, et al. Natural macrocyclic molecules have a possible limited structural diversity. Mol Divers. 2007;11: 115–118. doi:10.1007/s11030-007-9065-5

2. Hill TA, Shepherd NE, Diness F, Fairlie DP. Constraining Cyclic Peptides To Mimic Protein Structure Motifs. Angewandte Chemie International Edition. 2014;53: 13020–13041. doi:10.1002/anie.201401058

3. D'Souza VT, Lipkowitz KB. Cyclodextrins: Introduction. Chem Rev. 1998;98: 1741–1742. doi:10.1021/cr980027p

4. Palei S, Mootz HD. Preparation of Semisynthetic Peptides Macrocycles Using Split Inteins. Methods Mol Biol. 2017;1495: 77–92. doi:10.1007/978-1-4939-6451-2_6

5. Kwitkowski VE, Prowell TM, Ibrahim A, Farrell AT, Justice R, Mitchell SS, et al. FDA approval summary: temsirolimus as treatment for advanced renal cell carcinoma. Oncologist. 2010;15: 428–435. doi:10.1634/theoncologist.2009-0178

6. Raymond E, Alexandre J, Faivre S, Vera K, Materman E, Boni J, et al. Safety and Pharmacokinetics of Escalated Doses of Weekly Intravenous Infusion of CCI-779, a Novel mTOR Inhibitor, in Patients With Cancer. JCO. 2004;22: 2336–2347. doi:10.1200/JCO.2004.08.116

7. Goodin S. Novel cytotoxic agents: Epothilones. Am J Health Syst Pharm. 2008;65: S10–S15. doi:10.2146/ajhp080089

8. Goodin S. Ixabepilone: A novel microtubule-stabilizing agent for the treatment of metastatic breast cancer. Am J Health Syst Pharm. 2008;65: 2017–2026. doi:10.2146/ajhp070628

9. Stotani S, Giordanetto F. Overview of Macrocycles in Clinical Development and Clinically Used. Practical Medicinal Chemistry with Macrocycles. John Wiley & Sons, Ltd; 2017. pp. 411–499. doi:10.1002/9781119092599.ch16

10. Pedersen CJ. The Discovery of Crown Ethers. Science. 1988;241: 536–540. doi:10.1126/science.241.4865.536

11. Batten SR, Robson R. Catenane and Rotaxane Motifs in Interpenetrating and Self-Penetrating Coordination Polymers. Molecular Catenanes, Rotaxanes and Knots. John Wiley & Sons, Ltd; 2007. pp. 77–106. doi:10.1002/9783527613724.ch05

12. Yudin AK. Macrocycles: lessons from the distant past, recent developments, and future directions. Chem Sci. 2014;6: 30–49. doi:10.1039/C4SC03089C

13. Marsault E, Peterson ML. Macrocycles are great cycles: applications, opportunities, and challenges of synthetic macrocycles in drug discovery. J Med Chem. 2011;54: 1961–2004. doi:10.1021/jm1012374

14. Driggers EM, Hale SP, Lee J, Terrett NK. The exploration of macrocycles for drug discovery--an underexploited structural class. Nat Rev Drug Discov. 2008;7: 608–624. doi:10.1038/nrd2590

15. Mallinson J, Collins I. Macrocycles in new drug discovery. Future Medicinal Chemistry. 2012;4: 1409–1438. doi:10.4155/fmc.12.93

16. Dougherty PG, Qian Z, Pei D. Macrocycles as protein-protein interaction inhibitors. Biochem J. 2017;474: 1109–1125. doi:10.1042/BCJ20160619

17. Bell IM, Gallicchio SN, Abrams M, Beese LS, Beshore DC, Bhimnathwala H, et al. 3-Aminopyrrolidinone Farnesyltransferase Inhibitors: Design of Macrocyclic Compounds with Improved

Pharmacokinetics and Excellent Cell Potency. J Med Chem. 2002;45: 2388–2409. doi:10.1021/jm010531d

18. Leung SSF, Sindhikara D, Jacobson MP. Simple Predictive Models of Passive Membrane Permeability Incorporating Size-Dependent Membrane-Water Partition. J Chem Inf Model. 2016;56: 924–929. doi:10.1021/acs.jcim.6b00005

19. Leung SSF, Mijalkovic J, Borrelli K, Jacobson MP. Testing physical models of passive membrane permeation. J Chem Inf Model. 2012;52: 1621–1636. doi:10.1021/ci200583t

20. Rezai T, Bock JE, Zhou MV, Kalyanaraman C, Lokey RS, Jacobson MP. 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: 14073–14080. doi:10.1021/ja063076p

21. Giordanetto F, Kihlberg J. Macrocyclic Drugs and Clinical Candidates: What Can Medicinal Chemists Learn from Their Properties? J Med Chem. 2014;57: 278–295. doi:10.1021/jm400887j

22. Dömling A. Small molecular weight protein-protein interaction antagonists: an insurmountable challenge? Curr Opin Chem Biol. 2008;12: 281–291. doi:10.1016/j.cbpa.2008.04.603

23. Doak BC, Over B, Giordanetto F, Kihlberg J. Oral druggable space beyond the rule of 5: insights from drugs and clinical candidates. Chem Biol. 2014;21: 1115–1142. doi:10.1016/j.chembiol.2014.08.013

24. Villar EA, Beglov D, Chennamadhavuni S, Porco JA, Kozakov D, Vajda S, et al. How proteins bind macrocycles. Nat Chem Biol. 2014;10: 723–731. doi:10.1038/nchembio.1584

25. Beck B, Larbig G, Mejat B, Magnin-Lachaux M, Picard A, Herdtweck E, et al. Short and Diverse Route Toward Complex Natural Product-Like Macrocycles. Org Lett. 2003;5: 1047–1050. doi:10.1021/ol034077e

26. Liao GP, Abdelraheem EMM, Neochoritis CG, Kurpiewska K, Kalinowska-Tłuścik J, McGowan DC, et al. Versatile Multicomponent Reaction Macrocycle Synthesis Using α-Isocyano-ω-carboxylic Acids. Org Lett. 2015;17: 4980–4983. doi:10.1021/acs.orglett.5b02419

27. Madhavachary R, Abdelraheem EMM, Rossetti A, Twarda-Clapa A, Musielak B, Kurpiewska K, et al. Two-Step Synthesis of Complex Artificial Macrocyclic Compounds. Angew Chem Int Ed Engl. 2017;56: 10725–10729. doi:10.1002/anie.201704426

28. Vishwanatha TM, Bergamaschi E, Dömling A. Sulfur-Switch Ugi Reaction for Macrocyclic Disulfide-Bridged Peptidomimetics. Org Lett. 2017;19: 3195–3198. doi:10.1021/acs.orglett.7b01324

29. Abdelraheem EMM, Shaabani S, Dömling A. Artificial Macrocycles. Synlett. 2018;29: 1136–1151. doi:10.1055/s-0036-1591975

30. Wang W, Groves MR, Dömling A. Artificial Macrocycles as IL-17A/IL-17RA Antagonists. Medchemcomm. 2018;9: 22–26. doi:10.1039/C7MD00464H

31. Magiera-Mularz K, Skalniak L, Zak KM, Musielak B, Rudzinska-Szostak E, Berlicki Ł, et al. Bioactive Macrocyclic Inhibitors of the PD-1/PD-L1 Immune Checkpoint. Angewandte Chemie International Edition. 2017;56: 13732–13735. doi:10.1002/anie.201707707

32. Neochoritis CG, Kazemi Miraki M, Abdelraheem EMM, Surmiak E, Zarganes-Tzitzikas T, Łabuzek B, et al. Design of indole- and MCR-based macrocycles as p53-MDM2 antagonists. Beilstein J Org Chem. 2019;15: 513–520. doi:10.3762/bjoc.15.45

33. Estrada-Ortiz N, Neochoritis CG, Twarda-Clapa A, Musielak B, Holak TA, Dömling A. Artificial Macrocycles as Potent p53–MDM2 Inhibitors. ACS Med Chem Lett. 2017;8: 1025–1030. doi:10.1021/acsmedchemlett.7b00219

34. Kaserer T, Beck KR, Akram M, Odermatt A, Schuster D. Pharmacophore Models and Pharmacophore-Based Virtual Screening: Concepts and Applications Exemplified on Hydroxysteroid Dehydrogenases. Molecules. 2015;20: 22799–22832. doi:10.3390/molecules201219880

35. Spellmeyer DC, Wong AK, Bower MJ, Blaney JM. Conformational analysis using distance geometry methods. J Mol Graph Model. 1997;15: 18–36. doi:10.1016/s1093-3263(97)00014-4

36. Coutsias EA, Lexa KW, Wester MJ, Pollock SN, Jacobson MP. Exhaustive Conformational Sampling of Complex Fused Ring Macrocycles Using Inverse Kinematics. J Chem Theory Comput. 2016;12: 4674–4687. doi:10.1021/acs.jctc.6b00250

37. Vainio MJ, Johnson MS. Generating Conformer Ensembles Using a Multiobjective Genetic Algorithm. J Chem Inf Model. 2007;47: 2462–2474. doi:10.1021/ci6005646

38. Labute P. LowModeMD—Implicit Low-Mode Velocity Filtering Applied to Conformational Search of Macrocycles and Protein Loops. J Chem Inf Model. 2010;50: 792–800. doi:10.1021/ci900508k

39. Watts KS, Dalal P, Tebben AJ, Cheney DL, Shelley JC. Macrocycle conformational sampling with MacroModel. J Chem Inf Model. 2014;54: 2680–2696. doi:10.1021/ci5001696

40. Olanders G, Alogheli H, Brandt P, Karlén A. Conformational analysis of macrocycles: comparing general and specialized methods. J Comput Aided Mol Des. 2020 [cited 16 Feb 2020]. doi:10.1007/s10822-020-00277-2

41. Vulis M. Ring structures and the discrete Fourier transform. Advances in Applied Mathematics. 1985;6: 350–372. doi:10.1016/0196-8858(85)90016-8

42. Gerber P, Gubernator K, Müller K. Generic shapes for the conformation analysis of macrocyclic structures. Helvetica Chimica Acta. 2004;71: 1429–1441. doi:10.1002/hlca.19880710607

43. Halgren TA. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. Journal of Computational Chemistry. 1996;17: 490–519. doi:10.1002/(SICI)1096-987X(199604)17:5/6<490::AID-JCC1>3.0.CO;2-P

44. Rappe AK, Casewit CJ, Colwell KS, Goddard WA, Skiff WM. UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations. J Am Chem Soc. 1992;114: 10024–10035. doi:10.1021/ja00051a040

45. Whitty A, Zhong M, Viarengo L, Beglov D, Hall DR, Vajda S. Quantifying the chameleonic properties of macrocycles and other high-molecular-weight drugs. Drug Discov Today. 2016;21: 712–717. doi:10.1016/j.drudis.2016.02.005

46. Friedrich N-O, Flachsenberg F, Meyder A, Sommer K, Kirchmair J, Rarey M. Conformator: A Novel Method for the Generation of Conformer Ensembles. J Chem Inf Model. 2019;59: 731–742. doi:10.1021/acs.jcim.8b00704

47. Wang S, Witek J, Landrum GA, Riniker S. Improving Conformer Generation for Small Rings and Macrocycles Based on Distance Geometry and Experimental Torsional-Angle Preferences. J Chem Inf Model. 2020 [cited 14 Apr 2020]. doi:10.1021/acs.jcim.0c00025

48. Taylor R, Cole J, Korb O, McCabe P. Knowledge-based libraries for predicting the geometric preferences of druglike molecules. J Chem Inf Model. 2014;54: 2500–2514. doi:10.1021/ci500358p

49. Sindhikara D, Spronk SA, Day T, Borrelli K, Cheney DL, Posy SL. Improving Accuracy, Diversity, and Speed with Prime Macrocycle Conformational Sampling. J Chem Inf Model. 2017;57: 1881–1894. doi:10.1021/acs.jcim.7b00052

50. Groom CR, Bruno IJ, Lightfoot MP, Ward SC. The Cambridge Structural Database. Acta Cryst B. 2016;72: 171–179. doi:10.1107/S2052520616003954

51. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic Acids Res. 2000;28: 235–242. doi:10.1093/nar/28.1.235

52. Gerber PR. Topological Pharmacophore Description of Chemical Structures using MAB-Force-Field-Derived Data and Corresponding Similarity Measures. In: Carbó-Dorca R, Gironés X, Mezey PG, editors. Fundamentals of Molecular Similarity. Boston, MA: Springer US; 2001. pp. 67–81. Available: https://doi.org/10.1007/978-1-4757-3273-3_5

53. Gerber PR, Müller K. MAB, a generally applicable molecular force field for structure modelling in medicinal chemistry. J Comput Aided Mol Des. 1995. doi:10.1007/BF00124456

54. Cole JC, Korb O, McCabe P, Read MG, Taylor R. Knowledge-Based Conformer Generation Using the Cambridge Structural Database. J Chem Inf Model. 2018;58: 615–629. doi:10.1021/acs.jcim.7b00697

55. Kirchmair J, Markt P, Distinto S, Wolber G, Langer T. Evaluation of the performance of 3D virtual screening protocols: RMSD comparisons, enrichment assessments, and decoy selection--what can we learn from earlier mistakes? J Comput Aided Mol Des. 2008;22: 213–228. doi:10.1007/s10822-007-9163-6

56. Friedrich N-O, de Bruyn Kops C, Flachsenberg F, Sommer K, Rarey M, Kirchmair J. Benchmarking Commercial Conformer Ensemble Generators. J Chem Inf Model. 2017;57: 2719–2728. doi:10.1021/acs.jcim.7b00505

57. Bai F, Liu X, Li J, Zhang H, Jiang H, Wang X, et al. Bioactive conformational generation of small molecules: A comparative analysis between force-field and multiple empirical criteria based methods. BMC Bioinformatics. 2010;11: 545. doi:10.1186/1471-2105-11-545

58. Schulz-Gasch T, Schärfer C, Guba W, Rarey M. TFD: Torsion Fingerprints as a new measure to compare small molecule conformations. J Chem Inf Model. 2012;52: 1499–1512. doi:10.1021/ci2002318

59. Todeschini R. Molecular descriptors. Recent Advances in QSAR Studies Springer. 2020 [cited 2 Jan 2020]. Available: https://www.academia.edu/2884066/Molecular_descriptors

60. Hunter JD. Matplotlib: A 2D Graphics Environment. Computing in Science Engineering. 2007;9: 90–95. doi:10.1109/MCSE.2007.55

61. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nature Methods. 2020;17: 261–272. doi:10.1038/s41592-019-0686-2

62. Hendrickson JB. Molecular Geometry. IV. The Medium Rings. J Am Chem Soc. 1964;86: 4854–4866. doi:10.1021/ja01076a027

63. Engler EM, Andose JD, Schleyer PVR. Critical evaluation of molecular mechanics. J Am Chem Soc. 1973;95: 8005–8025. doi:10.1021/ja00805a012

64. Hagelueken G, Albrecht SC, Steinmetz H, Jansen R, Heinz DW, Kalesse M, et al. The Absolute Configuration of Rhizopodin and Its Inhibition of Actin Polymerization by Dimerization. Angewandte Chemie International Edition. 2009;48: 595–598. doi:10.1002/anie.200802915

65. Kuhn B, Mohr P, Stahl M. Intramolecular Hydrogen Bonding in Medicinal Chemistry. J Med Chem. 2010;53: 2601–2611. doi:10.1021/jm100087s

66. Alex A, Millan DS, Perez M, Wakenhut F, Whitlock GA. Intramolecular hydrogen bonding to improve membrane permeability and absorption in beyond rule of five chemical space. Med Chem Commun. 2011;2: 669–674. doi:10.1039/C1MD00093D

67. Danelius E, Poongavanam V, Peintner S, Wieske LHE, Erdélyi M, Kihlberg J. Solution Conformations Explain the Chameleonic Behaviour of Macrocyclic Drugs. Chemistry – A European Journal. 2020;26: 5231–5244. doi:10.1002/chem.201905599

68. Rossi Sebastiano M, Doak BC, Backlund M, Poongavanam V, Over B, Ermondi G, et al. Impact of Dynamically Exposed Polarity on Permeability and Solubility of Chameleonic Drugs Beyond the Rule of 5. J Med Chem. 2018;61: 4189–4202. doi:10.1021/acs.jmedchem.8b00347

69. Lammers M, Neumann H, Chin JW, James LC. Acetylation regulates Cyclophilin A catalysis, immunosuppression and HIV isomerization. Nat Chem Biol. 2010;6: 331–337. doi:10.1038/nchembio.342

70. Northfield SE, Wielens J, Headey SJ, Williams-Noonan BJ, Mulcair M, Scanlon MJ, et al. Cyclic Hexapeptide Mimics of the LEDGF Integrase Recognition Loop in Complex with HIV-1 Integrase. ChemMedChem. 2018;13: 1555–1565. doi:10.1002/cmdc.201800129

71. Pía E, Toba R, Chas M, Peinador C, Quintela JM. Synthesis of new viologen macrocycles with intramolecular charge transfer. Tetrahedron Letters. 2006;47: 1953–1956. doi:10.1016/j.tetlet.2006.01.073

72. Kamenik AS, Kraml J, Hofer F, Waibl F, Quoika PK, Kahler U, et al. Macrocycle Cell Permeability Measured by Solvation Free Energies in Polar and Apolar Environments. J Chem Inf Model. 2020;60: 3508–3517. doi:10.1021/acs.jcim.0c00280

73. Kamenik AS, Lessel U, Fuchs JE, Fox T, Liedl KR. Peptidic Macrocycles - Conformational Sampling and Thermodynamic Characterization. J Chem Inf Model. 2018;58: 982–992. doi:10.1021/acs.jcim.8b00097

74. Cottrell SJ, Olsson TSG, Taylor R, Cole JC, Liebeschuetz JW. Validating and Understanding Ring Conformations Using Small Molecule Crystallographic Data. J Chem Inf Model. 2012;52: 956–962. doi:10.1021/ci200439d

75. Bruno IJ, Cole JC, Kessler M, Luo J, Motherwell WDS, Purkis LH, et al. Retrieval of Crystallographically-Derived Molecular Geometry Information. J Chem Inf Comput Sci. 2004;44: 2133–2144. doi:10.1021/ci049780b

76. Cleves AE, Jain AN. ForceGen 3D structure and conformer generation: from small lead-like molecules to macrocyclic drugs. J Comput Aided Mol Des. 2017;31: 419–439. doi:10.1007/s10822-017-0015-8

Chapter 4

Reverse Docking for the Identification of Molecular Targets of Anticancer Compounds

Angel Jonathan Ruiz-Moreno, Alexander Dömling, and Marco Antonio Velasco-Velázquez

This chapter has been published in Robles-Flores M. (eds) Cancer Cell Signaling. Methods in Molecular Biology, vol 2174. Humana, New York, NY.

Abstract

Molecular docking is a useful and powerful computational method for the identification of potential interactions between small molecules and pharmacological targets. In reverse docking, the ability of one or a few compounds to bind a large dataset of proteins is evaluated *in silico*. This strategy is useful for identifying molecular targets of orphan bioactive compounds, proposing new molecular mechanisms, finding alternative indications of drugs, or predicting drug toxicity. Herein, we describe a detailed reverse docking protocol for the identification of potential targets for 4-hydroxycoumarin (4-HC). Our results showed that RAC1 is a target of 4-HC, which partially explains the biological activities of 4-HC on cancer cells. The strategy reported here can be easily applied to other compounds and protein datasets.

Introduction

Molecular docking was first described by Kuntz in 1982 [1]. To date, it has become a central tool in virtual drug screening, given its ability to predict ligand–target interaction. Molecular docking comprises two major tasks. First, the sampling algorithm predicts the many conformations that the ligand can assume within the pocket of interest (referred as poses). Then, a scoring function predicts the binding affinity between ligand and receptor for each pose. The generated binding poses are then ranked based on their binding affinity scores [2]. Ideally, the top-ranked pose should correspond to the ligand-binding mode present in nature. On the other hand, scoring functions are usually employed for ranking and filtering large databases of compounds in structure-based virtual screening. The highest-ranked ligands have the best binding affinity scores and, thus, can be considered lead compounds [3].

An additional application of molecular docking is the identification of potential molecular targets of orphan bioactive compounds; this strategy is called reverse docking. In contrast to the traditional molecular docking approach, reverse docking is used to identifying potential receptors for a given ligand among a large number of structures. Because of that, reverse docking can be used to discover targets for existing drugs, natural compounds, and novel molecules. Consequently, reverse docking allows the identification of the molecular mechanism of a substance with an unknown target, the finding of alternative indications of drugs (repurposing), or the prediction of adverse drug reactions or toxicity [4]. 4-Hydroxycoumarin (4-HC) has antimetastatic and antineoplastic activities in preclinical models of melanoma [5–7] (Fig. 1a). To identify the potential targets that mediate the reported effects of 4-HC, we performed reverse docking between 4-HC and a human protein dataset (Fig. 1b) retrieved from the Research Collaboratory for Structural Bioinformatics Protein Data Bank (rcsbPDB— https://www.rcsb.org/-). This example allows the explanation of the process of setting up a reverse docking experiment.

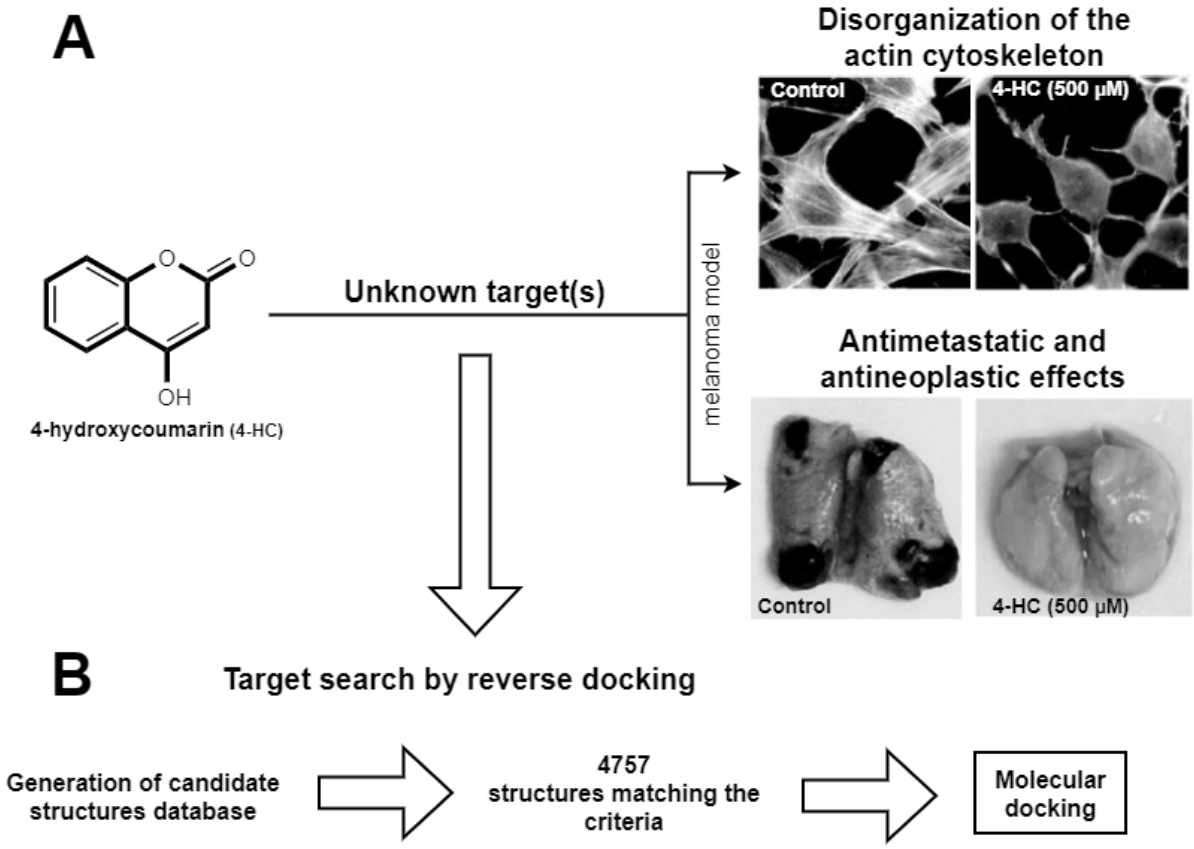


Figure 1. Reported effects of 4-HC and strategy for target identification. (a) 4-HC disorganizes the actin cytoskeleton in melanoma cells and has antimetastatic activity in a murine melanoma model. However, the target mediating these effects of 4-HC is unknown. (b) Proposed screening for identification of potential targets of 4-HC by reverse docking.

Materials

Computational Workstation

A computer with at least four available cores. For generation of data presented here, we used two Central Processor Units (CPUs) model Intel® Xeon® W3503 with two cores each (see Note 1), 12 GB of Random-access Memory (RAM), and 2 TB of Hard Drive (HD) (see Note 2) running Linux Ubuntu 16.04 Long Term Stable (LTS) for 64 bits architecture (see Note 3).

Python Environments Manager (See Note 4)

Miniconda3 for creation, management, and installation of python packages. It can be downloaded from https://docs.conda.io/en/ latest/miniconda.html.

Molecular Docking Software

GOLD from the Cambridge Structural Database suite 2019 (CSD suite 2019) of the Cambridge Crystallography Data Centre (CCDC) through the Python Application Programming Interface (Python API) on Python 2.7 (see Note 5).

Bioinformatics and Cheminformatic Tools

1. Conda packages of MDtraj 1.9.1 [8] and PBDFixer 1.5 [9] for automation of protein preparation. They are available at http://mdtraj.org/1.9.3/ and https://github.com/ pandegroup/pdbfixer.

2. Command line installation of Fpocket 3.0 [10] for the identification of protein pockets in the protein dataset. Fpocket can be downloaded from http://fpocket.sourceforge.net/.

3.The Graphic User Interface (GUI) of MarvinSketch and Standardizer for ligand drawing and preparation, available at ChemAxon (http://www.chemaxon.com).

4. Conda package of Pymol 2.3.1 (https://pymol.org/2/), available at https://anaconda.org/schrodinger/pymol.

5. Conda package of OpenBabel 2.4.1 [11], which can be downloaded from https://anaconda.org/openbabel/openbabel.

6. Protein-Ligand Interaction Profiler (PLIP) [12] server (https://projects.biotec.tu-dresden.de/plip-web/plip/index).

Analysis Tools Working on Python 3.6 1

1. Conda package of Scipy 1.3.0 [13], an open-source software for mathematics, science, and engineering available at https:// www.scipy.org/.

2. Conda package of The Macromolecular Transmission Format (MMTF) [14] downloadable at https://anaconda.org/condaforge/ mmtf-python.

3. Conda package of BioPython 1.72 [15], which contains freely available tools for biological computation. Available at https:// biopython.org/.

Methods

Generation of Protein Structures Database

The protein dataset of this example was built accordingly to the following advanced search parameters into rcsdPDB (see Note 6): – Experimental Method: X-ray. – Molecule: Protein. – Organism: Homo sapiens (only). – X-ray Resolution: 0–2.5 A ° . – Sequence identity: 90%. Our search retrieved 4757 biological assemblies of proteins that were downloaded as pdb format files to generate the protein dataset analyzed.

Setting Up the Software and Conda Environments

These procedures aim to install the tools that will be used in the protocol. At the end of these steps, all the software will be ready to start the ligand and protein preparations and the docking procedure.

1. Download and install Fpocket 3.0, MarvinSketch, and Standardizer according to the developer instructions.

2. Download and install the Miniconda3 installer for the proper platform.

3. Create two Python environments for docking and analysis, respectively by typing into the terminal:

(a) $conda create -n Docking python=2.7.

(b) $conda create -n Analysis python= 3.6.

For further info about environment activation and management of packages, see Note 4.

Inside of *Docking environment*, install the CSD suite 2019 from CCDC, including GOLD. Follow the developer instructions for installation. Then install Mdtraj 1.9.1 and PDBFixer 1.5. by typing into the terminal:

(c) $conda install -c omnia mdtraj = 1.9.1 pdbfixer = 1.5

Inside of the *Analysis environment*, install OpenBabel, Pymol, and all the tools listed on Analysis Tools Working on Python 3.6 1 subheading typing:

(d) $conda install -c openbabel openbabel.

(e) $conda install -c schrodinger pymol.

(f) $conda install scipy.

(g) $conda install -c conda-forge mmtf-python.

(h) $conda install -c anaconda biopython.

Protein Dataset Preparation for Reverse Docking Assays (See Note 7*)*

These steps will prepare each protein structure and store the corresponding information in new files. Once completed, the whole dataset of proteins will be ready for being employed in molecular docking. Figure 2a shows the differences in a protein from the dataset after preparation.

1. Activate and use the Analysis environment. Remove solvent molecules (water, ions, dimethyl sulfoxide, glycerol, etc.) and cocrystallized ligand(s) by using the *remove_solvent()* function of Mdtraj and *removeHeterogens()* of PDBfixer, respectively.

2. Complete the peptidic chain and replace nonstandard residues (i.e., selenomethionine (MSE) to methionine (MET)) by using the functions *findNonstandardResidues()* and *replaceNonstandardResidues()* from PDBFixer.

3. Find and add the missing residues and atoms with PDBfixer using *findMissingResidues()*, *findMissingAtoms()*, and *addMissingAtoms()*.

4. Add the missing hydrogens for protein structures at pH 7.4 utilizing the function and variable *addMissingHydrogens(7.4)* on PDBfixer.

5. Assign partial charges using *assign_partial_charges()* from the Protein module of the CSD suite.

6. Save the prepared protein in mol2 and pdb file formats by employing the *write()* function from the MoleculeWriter module of CSD suite (such files will be used later). For instance, files can be named pdbCode_prep.mol2 and pdbCode_prep.pdb, where pdbCode would be the PDB entry number of the structure in rcsbPDB.

Ligand Building and Preparation for Reverse Docking (See Note 8)

The goal of this procedure is obtaining the optimized 3D structure of the ligand(s) for docking experiments (i.e., including explicit hydrogens for pH 7.4, and the properly aromaticity perception).

1. Use MarvinSketch GUI to draw the ligand. The structure of 4-HC was generated from the SMILES code OC1=CC(=O)C2=CC=CC=C2O1. Alternatively, the option "Structure- >Name to Structure" can be used by typing 4-hydroxycoumarin.

2. On MarvinSketch, select the 4-HC and perform a pH-dependent protonation analysis utilizing "Calculations > Protonation > pKa" and the default settings. Figure 2b shows the optimized structure of 4-HC and the results obtained in protonation analysis. Note that the ionic form of 4-HC is prevalent at pH 7.4, whereas the percentage of the neutral species is almost zero. Because of that, we worked with the ionic formof 4-HC with SMILES code [O-]C1=CC(=O) OC2 = CC=CC=C12.

3. Using the SMILES code, save the ionic form of the 4-HC in a new file called 4-HC.mol2 using "File > Save As" on MarvinSketch.

4. Open Standardizer and use the 4-HC.mol2 file as input. Click "Next" and select the molecule standardization options "Add Explicit Hydrogens" and "Aromatize" (order is important). Click "Next" and save the file as 4-HC_standarized.mol2.

5. Activate the Analysis environment and use OpenBabel 2.4.1 to optimize the molecule using the MMFF94s forcefield by typing into the terminal: (a) obminimize -ff MMFF94s -sd -n 2500 -c 0.00001 -cut -rvdw 6.0 -rele 10.0 -pf 10 4-HC_standarized.mol2 > 4- HC_ready.mol2, where obminimize is the module of OpenBabel for energy minimization; -ff MMFF94s is the forcefield to use for optimization; -sd is the steepest descent algorithm;- n 2500 is the number of steps; -c 0.00001 is the convergence criteria; -cut enables the cutoff; -rvdw 6.0 is the cutoff for the Vander Walls distance; -rele 10.0 is the electrostatics cutoff; -pf 10 is the frequency to update the nonbonded pairs; and 4-HC_standarized.mol2 > 4- HC_ready.mol2 corresponds to the input and output file, respectively (see Note 9).

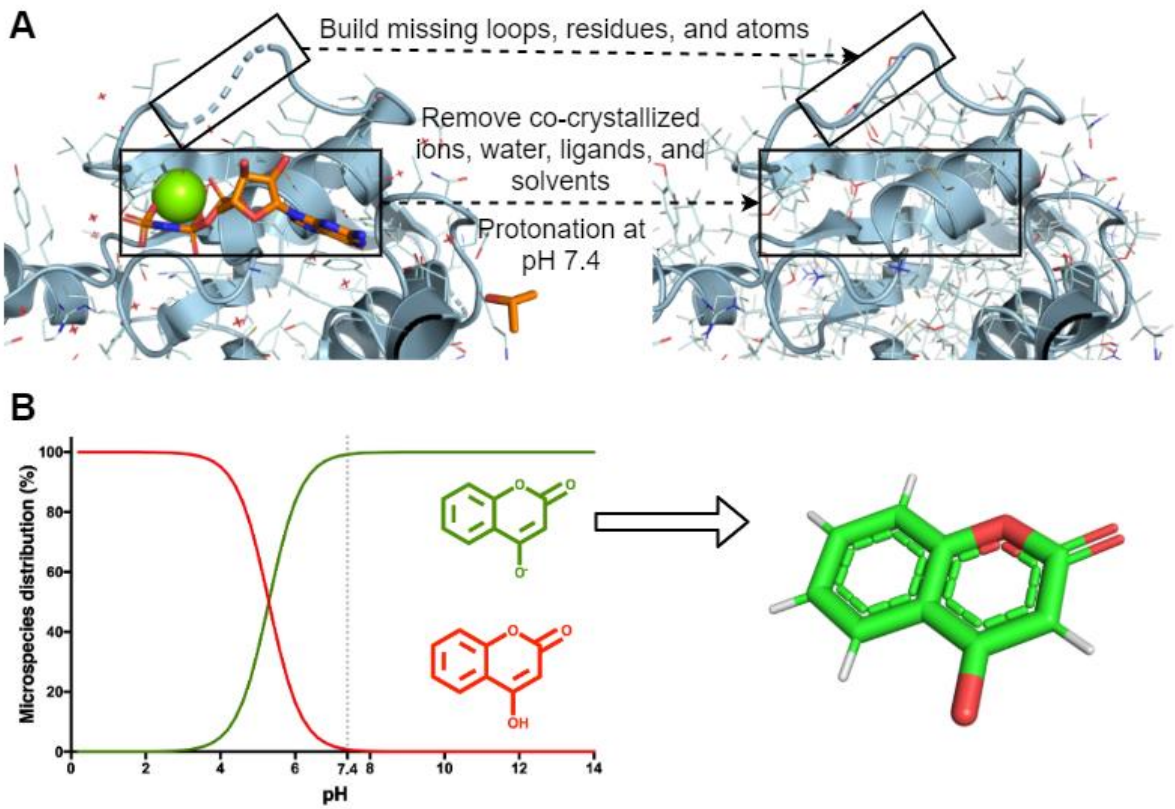


Figure 2. Proteins and ligand preparation for docking. (a) Structure of a model protein before and after preparation. Preparation must consider the removal of co-crystallized molecules (i.e., solvents), the building of missing loops, and the proper protonation state of the residues. (b) Preparation of 4-HC. Ligand preparation must include the analysis of the prevalence of the molecule at defined pH (green line) and selection of proper protonation and tautomeric state. The 3D energetically optimized geometry of 4-HC is shown on the right.

Protein Pocket Search for Reverse Docking

1. Utilize Fpocket 3.0 to find the pocket for each the pdbCode_prep. pdb file. Type in the terminal: $fpocket -F input_file.txt -v 500. Where input_file.txt is a text file in which each line contains the path to each pdbCode_prep.pdb file of the prepared protein dataset, and -v 500 is the number of Monte-Carlo iterations for the calculation of the pocket volume. The result of this run will be saved in a new folder for each protein-containing (among other things) the 3D coordinates of the pockets plus the protein structure in a pdb file and a txt file with several descriptors including the pocket volume (see Note 10).

2. Use Pymol inside of the Analysis environment (either using the GUI or by python programming) to extract the coordinates of pockets with a volume higher or equal to 150 Å 3 into a mol2 file for each protein. For instance, pdbCode_PocketNum.mol2. These files include all relevant pockets for one protein. Pockets smaller than 150 Å 3 are too small to allocate the 4-HC and, thus, are

irrelevant for this experiment. The generated mol2 files will be used as references for the reverse docking; therefore, their number corresponds to the number of dockings to perform. For this example, the number of references is approximately 50,000.

Reverse Docking

1. Activate the *Docking environment* to set up the reverse docking. Import the module Docker () from the CSD suite 2019 to initialize the GOLD docking engine.

2. Establish the following Docker() settings for the docking of the 4-HC into each protein using the reference coordinates of pockets:

(a) receptor =add_protein_file ('pdbCode_prep.mol2').

(b) reference=MoleculeReader ('pdbCode_pocketNum. mol2').

(c) BindingSiteFromPoint(receptor,reference.centre_of_- geometry(),6); where 6 is the number of Å to extend the pocket centre.

(d) Fitness_funtion = 'plp'; which means using PLPchemscore function for docking scoring.

(e) autoscale=10.0; recommended value for High Throughput Screening (HTS).

(f) add_ligand_file('4-HC_ready.mol2',10); where 10 refers to the number of best-scored poses requested as docking result.

(g) output_file = 'pdbCode_pocketNum.mol2'. For a clearer reference about the establishment of settings in a Pythonic way, see Fig. 3 (see Note 11).

3. Run the docking, which will generate a mol2 file for each pocket reference (pdbCode_pocketNum.mol2 from output_file variable). Such files contained the ten best-scored poses of 4-HC and thus were used for analysis.

```
from ccdc.docking import Docker
from ccdc.io import MoleculeReader

for 'pdbCode_prep.mol2' in 'prepProteins_Folder':
    for 'pdbCode_pocketNum.mol2' in 'pockets_Folder':
        if pdbCode '''in Protein''' == pdbCode '''in Pocket''':
            docker=Docker.settings
            receptor=settings.add_protein_file('pdbCode_prep.mol2')
            reference=MoleculeReader ('pdbCode_pocketNum.mol2')
            settings.BindingSiteFromPoint(receptor,reference.centre_of_geometry(),6.0)
            settings.fitness_function='plp'
            settings.autoscale=10.0
            settings.add_ligand_file('4-HC_ready.mol2',10)
            settings.output_directory='dockingResults_Folder'
            settings.output_file='pdbCode_pocketNum.mol2'
            docker.dock()
        else:
            continue
```

Figure 3. Python programming workflow for reverse docking using GOLD. The image shows a general script to establish the reverse docking settings through a python programming script in order to run the molecular docking using the Python API of GOLD.

Analysis of Docking Results

1. Active the Analysis environment containing the python libraries Scipy, MMTF, and BioPython.

2. Use the docking results to generate a table with the entry number of the protein and the highest score value (best docking result) for each evaluation.

3. Visualize such data in a probability distribution plot and select the potential targets among the proteins with a higher docking score. We selected as candidates 67 proteins with docking scores _55 (Fig. 4a). This cutoff value corresponds to the average docking score + 3.5 SD.

4. Generate a new table containing the entry name of the best docking results and their score. Use the MMTF python library to extract the sequence of each pdbCode using the function entity_list and add such data into a new column in the table.

5. For all entries into best docking results table, perform an all vs. all sequence alignment using the function alignment_- score = pairwise2.align.globalxx(target, reference, score_ only ¼ True) of BioPython library and calculate the identity percentage as follows:

$$\%Identity = \frac{Aligment\ score}{Length\ of\ reference}\ x\ 100$$

6. The result must be a matrix of size N _ N; where N is the number of entries in best docking results (67 for our example), and each data inside the matrix must be the % identity for each target vs. reference sequence alignment.

7. Analyze the matrix with the identity percentage values using the Single-Linkage Hierarchical Clustering (SLHC) algorithm to identify protein clusters from best docking results by using the functions inside of the Scipy module scipy.cluster.hierarchy. Then, sort the x and y axis of the matrix according to the hierarchical map and create a heatmap of the matrix. Further analysis may include a UniProt (https://www.uniprot.org/) search to identify Gene Ontology (GO) annotations for the best docking proteins.

Figure 4b shows the hierarchical clustering and identity heatmap for 4-HC potential targets. Clusters of proteins with high identity can be identified, suggesting that a common structure/domain could be mediating ligand–protein interaction. Our analysis identified a cluster of proteins with GO annotations that match the previously reported experimental evidence [5–7]. We focused on such group and selected the protein RAC1 (PDB code 4GZL), which participates in the control of the actin cytoskeleton organization as the best target candidate.

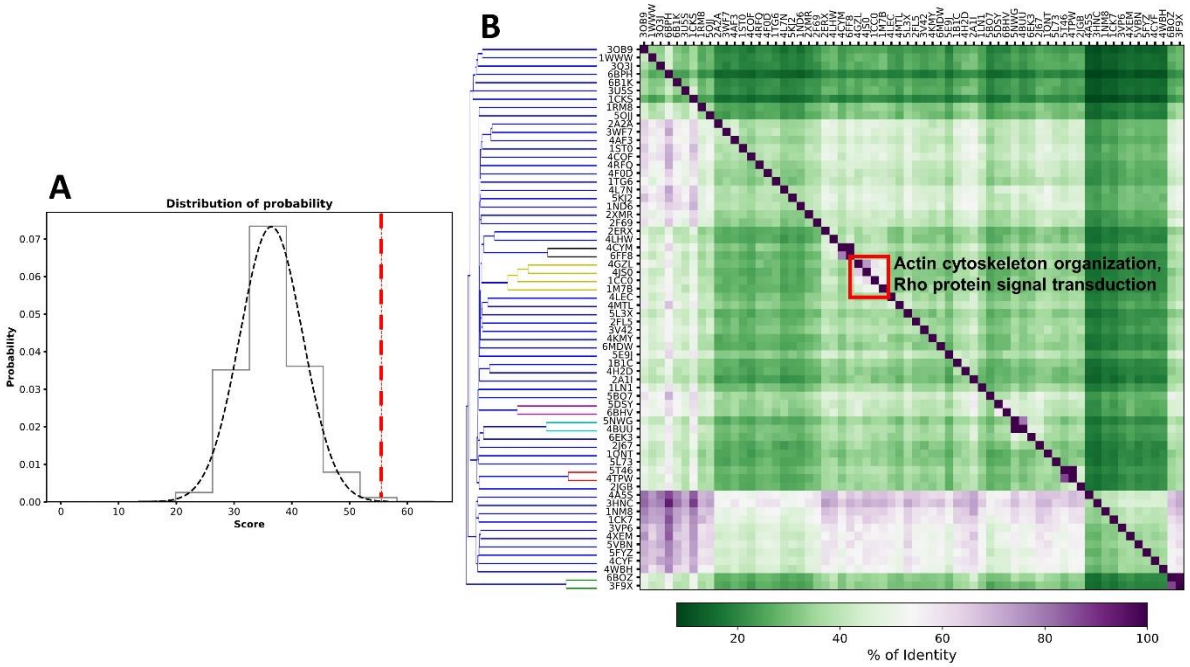


Figure 4. Distribution of probability and Single-Linkage Hierarchical Clustering of reverse docking results. (a) The best pose scores for each ligand–pocket pair showed a normal distribution. We considered as potential targets of 4-HC the proteins with docking scores _ mean + 3.5 SD (red line). (b) Potential targets were analyzed by SLHC based on their identity. We identified a protein cluster with actin cytoskeleton organization function (red square).

8. Generate a protein–ligand interaction map for the best docking pose of 4-HC in RAC1 (4GZL). For such purpose, use Pymol inside the Analysis environment to create a 4-HC-RAC1 complex using the 4GZL_1.mol2 file from dockingResults_Folder, and 4GZL_prep.mol2 file from

prepProteins_Folder, corresponding to docking poses of 4-HC and RAC1, respectively. Save the #1 pose and the protein into a Complex.pdb file. Use this file to create the interaction map by the PLIP server, listed on Subheading 2.4, item 6. The binding mode of 4-HC to RAC1 (4GZL) is shown if Fig. 5. This result generates a new hypothesis about the mechanism of action of 4-HC.

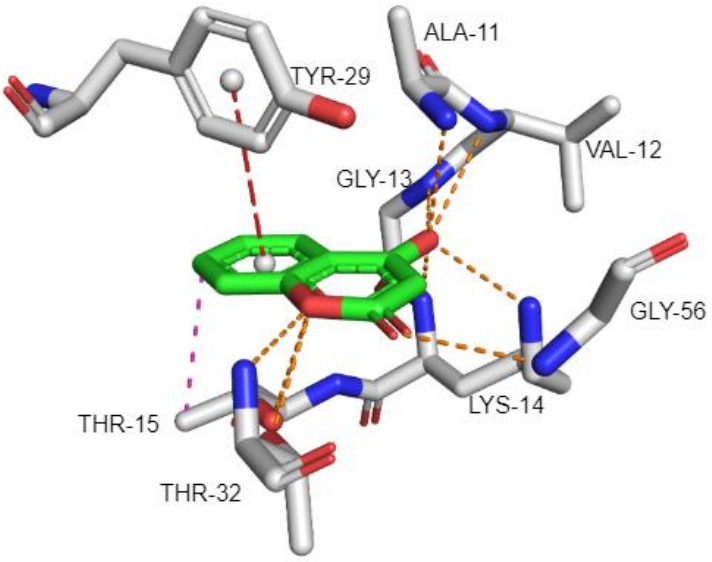


Figure 5. Interaction map of 4-HC with RAC1 (4GZL). Analysis of docking best results and SLHC indicated that RAC1 (white sticks) could be a potential target for 4-HC (green sticks) by forming several hydrogen bonds (orange dashed lines), a perpendicular π-staking (red dashed line), and hydrophobic interactions (magenta dashed line).

Notes

1. Most of docking algorithms and software are optimized for running even in low-specs computers. However, recent versions of docking programs, such as GOLD [2], support running in parallel through the Python API. Because of that, reverse docking experiments can be performed in multicore machines, as supercomputers or clusters.

2. The amount of data generated in virtual HTS experiments require enough available space into Hard Drive (HD) or Solid-state Drive (SSD). In our example, the total space of the workstation employed for running the example was 2 TB, and the space utilized for protein dataset, pockets, ligand, docking results, and analysis was 11.9 GB.

3. Several of the tools employed in this method have been developed for running into Windows, Mac OS, and Linux, but we strongly advise to use a Linux based OS (e.g., Ubuntu or Debian) because it improves software stability. We worked in Linux operative system Ubuntu 16.04.

4. Many of the tools required on this protocol will be managed by Miniconda3. Moreover, most of the tasks described in methods were automatized by using Python 2.7 and Python 3.6 programming. For more info about Miniconda3 visit https:// docs.conda.io/projects/conda/en/latest/, and for Python visit https://www.python.org/.

5. Further information about GOLD can be found at https:// www.ccdc.cam.ac.uk/solutions/csd-discovery/components/ gold/. We employed GOLD due to its versatility and accuracy [16]. However, this protocol can be adapted for running using another molecular docking software. For instance, running a reverse docking using AutoDock Vina can be achieved by Bourne-again shell (BASH) scripting on Linux OS or through the Graphic User Interface (GUI) [17]. The CSD suite containing GOLD can be employed either by GUI or using the Python API. For more info about available platforms, installation, and usage, visit https://downloads.ccdc.cam.ac.uk/documentation/API/installation_notes.html.

6. Protein dataset definition is a key task of reverse docking because it depends on the question aimed to address. For instance, the dataset should only include kinases if the hypothesis is that the compound (or series of compounds) can bind to kinases. A different dataset should be used for evaluating the ability of a compound to bind to proteins of a specific pathogenic microorganism. Independently of the characteristics that define the dataset, structural information of the proteins can be retrieved by searching into public databases as rcsbPDB or PDBe (https://www.ebi.ac.uk/pdbe/), both members of the Worldwide Protein Data Bank (wwPDB - https://www.wwpdb. org/-). Additional sources of structures available for reverse docking can be found in ref. 18 [18].

7. Alternatively, protein preparation can be achieved using GUIs as Dock Prep module of Chimera [19], which may be easier for users with less programming background. Whatever the approach selected for protein preparation; the steps are the same.

8. Ligand structures could come from various sources. This method describes the building of a ligand from zero. However, existing databases with molecules from different sources and chemical identity. For example, the ZINC15 database (https:// zinc15.docking.org/) which contains millions of molecular structures that can be used for *in silico* experiments.

9. Docking software cannot handle all of the file formats available to represent molecules. Thus, the user must select the proper file format for the docking software to be used. When working in GOLD, the recommended file format for proteins and ligands is mol2.

10. Identification and extraction of protein pockets allow the selection of druggable pockets. Fpocket can also search the pockets in the context of cocrystallized ligands. This option is useful

if only known binding sites will be analyzed. Please note that running docking of a ligand over all the protein surface is not recommended because most docking software were not created for such purpose.

11. Different docking software packages employ different sampling algorithms and score functions to find and evaluate ligand– protein interactions. The parameters established in Fig. 3 are the criteria to reproduce this experiment using GOLD, but it does not represent a python script. It must be considered just as an example to create an automatized python script for docking running. For other GOLD score functions and setting, visit (https://www.ccdc.cam.ac.uk/support-and resources/support/search?q=Scoring%20function). The use of other software than GOLD or different settings could rise to different results.

References

1. Kuntz ID, Blaney JM, Oatley SJ, Langridge R, Ferrin TE. A geometric approach to macromolecule-ligand interactions. Journal of Molecular Biology. 1982;161: 269–288. doi:10.1016/0022-2836(82)90153-X

2. Meng X-Y, Zhang H-X, Mezei M, Cui M. Molecular Docking: A powerful approach for structure-based drug discovery. Curr Comput Aided Drug Des. 2011;7: 146–157.

3. Phatak SS, Stephan CC, Cavasotto CN. High-throughput and in silico screenings in drug discovery. Expert Opinion on Drug Discovery. 2009;4: 947–959. doi:10.1517/17460440903190961

4. Lee A, Lee K, Kim D. Using reverse docking for target identification and its applications for drug discovery. Expert Opinion on Drug Discovery. 2016;11: 707–715. doi:10.1080/17460441.2016.1190706

5. Velasco-Velázquez MA, Agramonte-Hevia J, Barrera D, Jiménez-Orozco A, Garćıa-Mondragón MJ, Mendoza-Patiño N, et al. 4-Hydroxycoumarin disorganizes the actin cytoskeleton in B16–F10 melanoma cells but not in B82 fibroblasts, decreasing their adhesion to extracellular matrix proteins and motility. Cancer Letters. 2003;198: 179–186. doi:10.1016/S0304-3835(03)00333-1

6. Velasco-Velázquez MA, Salinas-Jazmín N, Mendoza-Patiño N, Mandoki JJ. Reduced paxillin expression contributes to the antimetastatic effect of 4-hydroxycoumarin on B16-F10 melanoma cells. Cancer Cell International. 2008;8: 8. doi:10.1186/1475-2867-8-8

7. Salinas-Jazmín N, De La Fuente M, Jaimez R, Pérez-Tapia M, Pérez-Torres A, Velasco-Velázquez MA. Antimetastatic, antineoplastic, and toxic effects of 4-hydroxycoumarin in a preclinical mouse melanoma model. Cancer Chemother Pharmacol. 2010;65: 931–940. doi:10.1007/s00280-009-1100-z

8. McGibbon RT, Beauchamp KA, Harrigan MP, Klein C, Swails JM, Hernández CX, et al. MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. Biophys J. 2015;109: 1528–1532. doi:10.1016/j.bpj.2015.08.015

9. Eastman P, Friedrichs MS, Chodera JD, Radmer RJ, Bruns CM, Ku JP, et al. OpenMM 4: A Reusable, Extensible, Hardware Independent Library for High Performance Molecular Simulation. J Chem Theory Comput. 2013;9: 461–469. doi:10.1021/ct300857j

10. Le Guilloux V, Schmidtke P, Tuffery P. Fpocket: An open source platform for ligand pocket detection. BMC Bioinformatics. 2009;10: 168. doi:10.1186/1471-2105-10-168

11. O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR. Open Babel: An open chemical toolbox. Journal of Cheminformatics. 2011;3: 33. doi:10.1186/1758-2946-3-33

12. Salentin S, Schreiber S, Haupt VJ, Adasme MF, Schroeder M. PLIP: fully automated protein–ligand interaction profiler. Nucleic Acids Res. 2015;43: W443–W447. doi:10.1093/nar/gkv315

13. Jones E, Oliphant T, Peterson P. SciPy: Open source scientific tools for Python. 2001. Available: https://scholar.google.com/scholar_lookup?title=SciPy: open source scientific tools for Python&author=&publication_year=2001

14. Bradley AR, Rose AS, Pavelka A, Valasatava Y, Duarte JM, Prlić A, et al. MMTF—An efficient file format for the transmission, visualization, and analysis of macromolecular structures. PLOS Computational Biology. 2017;13: e1005575. doi:10.1371/journal.pcbi.1005575

15. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics. 2009;25: 1422–1423. doi:10.1093/bioinformatics/btp163

16. Lee M, Kim D. Large-scale reverse docking profiles and their applications. BMC Bioinformatics. 2012;13: S6. doi:10.1186/1471-2105-13-S17-S6

17. Chen F, Wang Z, Wang C, Xu Q, Liang J, Xu X, et al. Application of reverse docking for target prediction of marine compounds with anti-tumor activity. J Mol Graph Model. 2017;77: 372–377. doi:10.1016/j.jmgm.2017.09.015

18. Xu X, Huang M, Zou X. Docking-based inverse virtual screening: methods, applications, and challenges. Biophys Rep. 2018;4: 1–16. doi:10.1007/s41048-017-0045-8

19. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, et al. UCSF Chimera--a visualization system for exploratory research and analysis. J Comput Chem. 2004;25: 1605–1612. doi:10.1002/jcc.20084

Chapter 5

Repurposing the HCV NS3–4A protease drug boceprevir as COVID-19 therapeutics

Rick Oerlemans†, Angel Jonathan Ruiz-Moreno†, Yingying Cong, Nilima Dinesh Kumar, Marco A. Velasco-Velazquez, Constantinos G. Neochoritis, Jolanda Smith, Fulvio Reggiori, Matthew R. Groves, and Alexander Dömling

This chapter has been published in RSC Med. Chem., 2021, **12**, 370-379.

† Equal contributors

Abstract

The rapid growth of COVID-19 cases is causing an increasing death toll and paralyzing the world economy. De novo drug discovery takes years to move from idea and/or pre-clinic to market, and it is not a short-term solution for the current SARS-CoV-2 pandemic. Drug repurposing is perhaps the only short-term solution, while vaccination is a middle-term solution. Here, we describe the discovery path of the HCV NS3–4A protease inhibitors boceprevir and telaprevir as SARS-CoV-2 main protease (3CLpro) inhibitors. Based on our hypothesis that α-ketoamide drugs can covalently bind to the active site cysteine of the SARS-CoV-2 3CLpro, we performed docking studies, enzyme inhibition and co-crystal structure analyses and finally established that boceprevir, but not telaprevir, inhibits replication of SARS-CoV-2 and mouse hepatitis virus (MHV), another coronavirus, in cell culture. Based on our studies, the HCV drug boceprevir deserves further attention as a repurposed drug for COVID-19 and potentially other coronaviral infections as well.

Introduction

Since emerging in Wuhan, China, in December 2019, the coronavirus (CoV) disease 2019 (COVID-19) epidemic caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has progressed rapidly into a pandemic [1]. COVID-19 is characterized by fever, cough, fatigue, shortness of breath, pneumonia, and other respiratory tract symptoms, and in many cases progresses to death [1–3]. As of August 30, 2020, there have been more than 25 million confirmed cases, and more than 830.000 deaths reported worldwide. Moreover, COVID-19 is making an immense negative impact on the world's economy and has become a huge societal burden [4].

SARS-CoV2 is an enveloped, non-segmented, positive-sense RNA virus that is part of the order Nidovirales, in the CoV virus family, which is broadly distributed in humans and other mammals [5,6]. SARS-CoV-2 is classified into the beta-CoV genera. Recent studies highlighted that SARS-CoV-2 genes share >80% nucleotide identity and >89% nucleotide similarity with SARS-CoV genes [7–9]. Upon cell entry, two polypeptides, pp1a and pp1ab, are produced by the host translation machinery directly from the CoV genome [10,11]. These two polypeptides self-cleave proteolytically into 11 and 16 individual non-structural (nsp) proteins, respectively, that are essential for viral replication [12]. CoV encode either two or three proteases that are involved in the self-cleavage of pp1a and pp1ab. They are the papain-like protease (PLpro), present with nsp3, and the 3C-like proteinase (3CLpro) or Mpro, localized in nsp5 [13]. Most CoV encode two PLpros within nsp3, except the gamma-CoV, SARS-CoV, Middle East respiratory syndrome coronavirus (MERS-CoV) and SARS-CoV-2 [13,14]. Importantly, 3CLpro plays a critical role in CoV replication and unlike structural/accessory protein-encoding genes, displays a considerable similarity between members of CoV, in particular beta-CoV [10,11]. Therefore, it is a promising target for the discovery and the development of a pan-anti-CoV inhibitor [6,15].

We investigated FDA-approved drugs with electrophile warheads for their potential to inhibit the SARS-CoV-2 proteases (Fig. 1A). Both proteases are cysteine proteases. The great majority of cysteine protease inhibitors function by a covalent mechanism where the nucleophilic sulfhydryl forms an (ir)reversible bond with an electrophilic warhead (α-ketoamide, for instance) of the inhibitor (Fig. 1B). Such covalent inhibitors have several advantages, including an increased ligand efficiency, overcoming competition with native ligands and less recurrent dosing due to sustained duration of action [16]. Thus, we focused our attention on subgroups of drugs, e.g.

electrophiles such as α-ketoamides and nitriles, which potentially can undergo a covalent modification of the active site cysteine of SARS-CoV-2 proteases. Our first finding was that nitrile containing gliptins are potential SARS-CoV-2 protease inhibitors [15]. Here, we report that α-ketoamide hepatitis C virus (HCV) protease inhibitors are also inhibitors of SARS-CoV-2 3CLpro (Fig. 1C). In particular, we present the finding of boceprevir and telaprevir as 3CLpro SARS-CoV-2 inhibitors, suggested by docking studies, supported by biochemical and co-crystal structure analyses, and finally provide evidence that boceprevir, but not telaprevir, inhibits viral replication in cellular assays assessing the replication of mouse hepatitis virus (MHV), also a beta-CoV, and SARS-CoV-2.

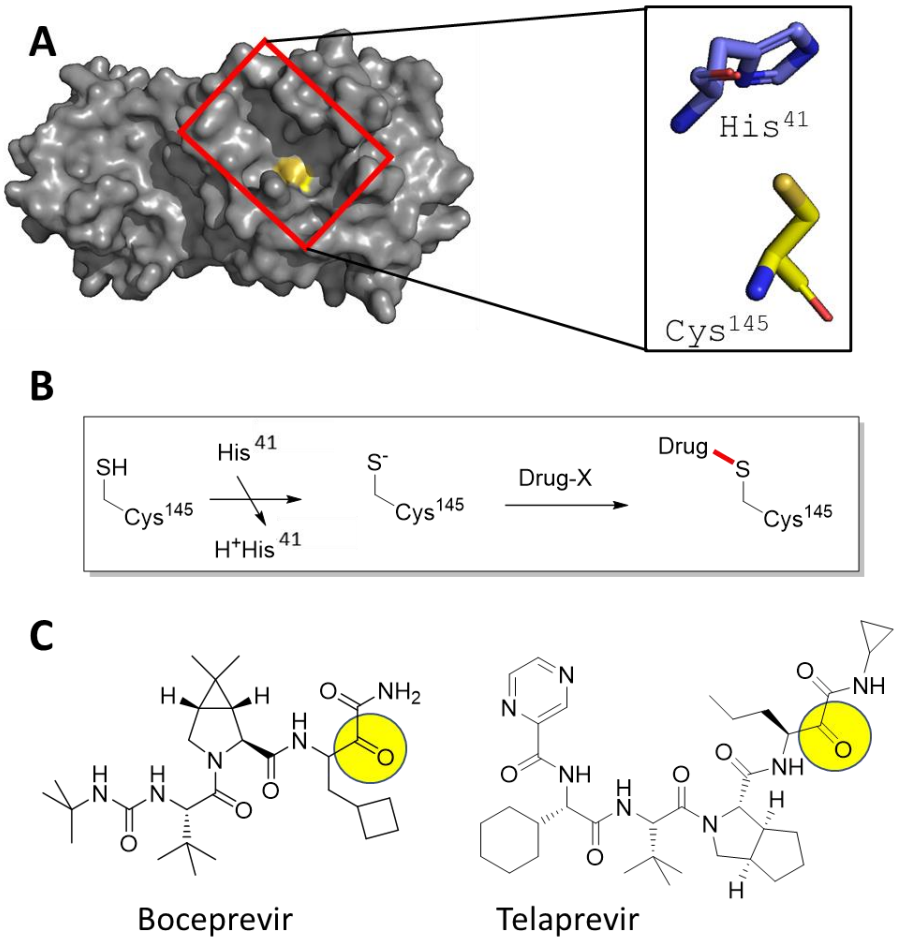


Figure 1. Hypothesis-driven illustration of α-ketoamide HCV NS3–4A protease inhibitors as covalent SARS-CoV-2 3CLpro inhibitors. A) Crystal structure of SARS-CoV-2 3CLpro (PDB ID 6LU7, grey surface presentation) highlighting the active site (red box) and the catalytic Cys145 (yellow). The catalytic dyad His41 and Cys145 are shown as stick presentation. B) Schematic mechanism of the addition of an electrophilic drug onto the active site Cys145 involving the His41 base. C) 2D structures of the marketed HCV drugs boceprevir and telaprevir, highlighting with a yellow circle the electrophilic α-ketoamide groups.

Methods

Chemicals and reagents

Boceprevir, telaprevir and bafilomycin A1 were obtained from Advanced Chemblocks Inc, Combi-Blocks and Sigma-Aldrich, respectively, and dissolved in dimethyl sulfoxide (DMSO, Sigma-Aldrich) at concentration of 100 mM, 100 mM and 200 µM, respectively. Boceprevir and telaprevir were stored at 4 °C, and bafilomycin at −20 °C.

Molecular docking

The methodology for covalent docking was performed as reported. Briefly, the crystal structure of SARS-CoV-2 3CLpro (PDB ID 6LU7) was used as a receptor, and boceprevir and telaprevir as ligands. Protein was prepared by removing co-crystallized waters, solvent molecules, and adding charges and hydrogens using Chimera 1.14 [17]. Isomeric SMILES codes for ligands were retrieved from PubChem and prepared for docking by setting the absolute stereo flags, adding explicit hydrogens and tautomeric states at pH 7.4. 3D coordinates were generated with Standardized 19.20.0 (http://www.chemaxon.com). For covalent docking, the sulfur atom of the reactive Cys145 of 3CLpro was used as the linker to form the covalent bond. For each ligand, 50 runs of genetic algorithm for the conformational search were performed and each pose was evaluated employing the PLP Chemscore scoring function using the GOLD software from the Cambridge Crystallography Data Center (CCDC) [18].

Cloning, protein expression and purification

An *E. coli* codon-optimized gene encoding SARS-CoV-2 3CLpro was purchased from Eurofins Genomics (Fig. S1). The gene was subcloned into the pET-28a(+) vector using NcoI and XhoI restriction enzymes (NEB). Initial tests showed that the resulting C-terminal 6His-tag was reducing enzymatic activity so in order to be able to obtain a native C terminus, PCR was performed with specific primers GCA GGT CTC GAG AGG CCC CTG AAA CGT AAC GCC GC (5′ → 3′) and CGC AAG CCC ATG GCG GC (5′ → 3′) to introduce a human Rhinovirus 3C protease (HRV 3Cpro) compatible cleavage site on the C terminus (SGVTFQGP). The PCR product was then digested and cloned into the pET-28a(+) vector using NcoI and XhoI.

The resulting pET-28a-3CLpro plasmid was transformed into competent BL21 Star (DE3) *E. coli* strain and a single colony was used to inoculate 20 ml of Luria–Bertani (LB) medium, supplemented with 50 µg/ml kanamycin and 35 µg/m chloramphenicol, before growth at 37 °C. After 16 h, the preculture was added to 2 l of LB medium (supplemented with 50 µg/ml kanamycin and 35 µg/ml chloramphenicol) and incubated at 37 °C in a shaking incubator (180 rpm). Expression of the fusion protein was induced by addition of 0.5 mM isopropyl-1-β;-D-thiogalactopyranoside (abcr chemicals) when the culture OD600 reached 0.6. At this point, the culture was transferred to an 18 °C shaking incubator (180 rpm) and after an overnight incubation, the bacteria were harvested by centrifugation at 4000g for 20 min. Pellets were resuspended in lysis buffer (50 mM Tris-HCl, pH 8, 300 mM NaCl, 10 mM imidazole, 20 µg/ml DNAse 1, 0.4 mg/ml lysozyme), lysed by sonication (50% amplitude, 2 s on/15 s off, for 5 min) and then clarified by centrifugation at 50 000g for 45 min. The supernatant was loaded on a 5 ml HiTrap HP column (GE Healthcare), washed with 5 column volumes of washing buffer (50 mM Tris-HCl, pH 8, 300 mM NaCl, 25 mM imidazole) and subsequently eluted in the elution buffer (50 mM Tris-HCl, pH 8, 300 mM NaCl, 250 mM imidazole). His-tagged HRV 3Cpro was added to the eluted fraction (1 : 10 w/w) and this mixture was dialyzed overnight at 4 °C against 50 mM Tris-HCl, 1 mM TCEP, 300 mM NaCl, to remove the imidazole and cleave the C-terminal 6xHis tag. Reverse nickel-NTA purification was then performed to elute untagged SARS-CoV-2 3CLpro, which was then further purified by size-exclusion chromatography (SEC) using a HiLoad 16/60 S200 (GE Healthcare) equilibrated with the SEC buffer (20 mM HEPES, pH 7.5, 50 mM NaCl). Finally, the purified SARS-CoV-2 3CLpro was concentrated to 5 mg/ml in a Vivaspin centrifugal concentrator (molecular weight cut-off (MWCO) 10 kDa, Sartorius).

Crystallization and structure determination

Initial apo seed crystals were obtained by the sitting drop vapour diffusion method at 18 °C with drops consisting of 0.5 µl protein solution (5 mg/ml) and 0.5 µl reservoir solution (0.2 M potassium sodium tartrate tetrahydrate, 0.2 M lithium acetate, 20% PEG 3350). These initial SARS-CoV-2 3CLpro crystals were used to make a seed stock using the Glass Seed Bead™ kit (Hampton). To obtain co-crystals of SARS-CoV-2 3CLpro with inhibitors, a pre-incubation co-crystallization protocol was followed. Briefly, protein at 0.5 mg/ml was incubated with 300 µM compound at 4 °C overnight. After centrifugation at 20 000g for 10 min to remove aggregates and precipitates, the protein solution was concentrated in a Vivaspin centrifugal concentrator (MWCO 10 kDa, Sartorius) to 5 mg/ml and an additional 300 µM compound was added and left to incubate at 4 °C for 3 h. After centrifugation, the pre-incubated protein solutions were used to obtain co-crystals

by the sitting drop vapour diffusion method [19]. For boceprevir the best crystals were grown in drops containing 1.5 µl pre-incubated protein solution (5 mg/ml), 1 µl reservoir solution (100 mM MES pH 6.75, 16% (w/v) PEG 6000, 5% (v/v) DMSO, 300 µM boceprevir) and 0.5 µl seed stock (diluted 1 : 500). For telaprevir, the best crystals were grown in drops containing 1.5 µl pre-incubated protein solution (5 mg/ml), 1 µl reservoir solution (100 mM MES pH 6.75, 18% (w/v) PEG 6000, 5% (v/v) DMSO, 300 µM telaprevir) and 0.5 µl seed stock (diluted 1 : 500). The cryo-protectant solution consisted of reservoir solution supplemented with 25% (v/v) glycerol.

X-ray diffraction data was collected on beamline P11 of the Deutsches Elektronen-Synchrotron (DESY) (Hamburg, Germany) at 100 K. Data integration and scaling was performed using XDS21 and Aimless [20] from the CCP4 software suite [21]. The DIMPLE software in CCP4 was used to obtain the solved structures, utilizing a previously solved SARS-CoV-2 3CLpro structure (ligand stripped) as the reference model (PDB: 6LU7). The DIMPLE output models were then subjected to iterative cycles of model building with COOT and refinement with REFMAC [22,23]. The structures were deposited into the PDB (6ZRT and 6ZRU for telaprevir and boceprevir, respectively). Data collection and refinement statistics can be found in Table S1. Similar structures were deposited by other groups (6XQS, 7C7P, 6WNP) during the course of this project.

3CLpro enzymatic assay

The *in vitro* 3CLpro enzymatic assay to assess compound inhibition was setup following a published protocol for SARS-CoV 3CLpro [24], with slight modifications. Briefly, a continuous kinetic FRET assay was used to measure SARS-CoV-2 3CLpro activity, against the substrate 2-aminobenzoyl-SVTLQSG-Tyr (NO2)-R (Genscript). The cleavage of the FRET substrate was monitored by measuring the increase in fluorescence in a FLUOstar Omega platereader (BMG labtech) at excitation and emission wavelengths of 330 nm and 420 nm, respectively. Total volume for the assay was 200 µl, containing 250 nM SARS-CoV-2 3CLpro and 100 µM substrate in the assay buffer (100 mM potassium phosphate, pH 8, 3% DMSO). The reaction was monitored for 15 min immediately after adding the FRET substrate. Initial rates were calculated via linear regression, using the first 2 min of the reaction. To determine the inhibitory activity of the compounds, the enzyme was incubated with 12 different concentrations of compound (20 nM – 400 µM telaprevir, 5 nM – 100 µM boceprevir) for 1 h at 25 °C, followed by initial rate determination. Dose response curves and IC50 values were obtained using the Prism8 software

by plotting the initial rates against inhibitor concentration and performing non-linear regression. The initial rates were normalized to controls, control without enzyme (100% inhibition) and control without compound (0% inhibition). All measurements were performed in triplicate.

Cell culture and the cytotoxic assay

The African green monkey Vero E6 cell line (ATCC CRL-1586), kindly provided by Gorben Pijlman (Wageningen University, Wageningen, the Netherlands) and the mouse LR7 cell line were maintained in Dulbecco's minimal essential medium (DMEM) (Gibco), high glucose supplemented with 10% fetal bovine serum (FBS) (Lonza), 100 U/ml penicillin, and 100 U/ml streptomycin. Cells were mycoplasma negative and maintained at 37 °C under 5% CO_2.

For the cytotoxic assay, LR7 or Vero E6 cells were seeded in 96-well plates at a density of 1 × 10^4 cells per well and cultured in DMEM containing 10% FBS at 37 °C under 5% CO2 for 24 h, followed by addition of serial dilutions (0–500 µM) of the tested drugs. Cells were allowed to grow for 8 h at 37 °C and proliferation was analyzed using the 3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide (MTT) method. Briefly, the medium was removed before to add 100 µl of 0.5 mg/ml MTT solution (Sigma-Aldrich) and incubate cells at 37 °C for 4 h. Then 100 µl of DMSO were added and incubated for 30 min to solubilize the formazan crystals. Absorbance of each well at OD490 was measured using a GloMax-Multi Detection System (Promega) and cell survival percentage was calculated as OD490 of the sample/OD490 of the control.

Virus stocks and titration

MHV strain A59 was propagated in LR7 cells in DMEM and the virus titer was determined on LR7 cells and the culture infectious dose (TCID50) units per ml of supernatant were calculated according to the Reed–Muench method [25]. The SARS-CoV-2 strain/NL/2020 was obtained from European Virus archive global (EVAg-010V-03903). For virus production, Vero E6 cells were infected at a multiplicity of infection (MOI) of 1 and 48 h post infection (p.i.) supernatants containing progeny virions were harvested, centrifuged, aliquoted and stored at −80 °C. The virus was passaged twice after receiving it from EVAg and passage 2 virus was used for subsequent experiments. Titration was performed using plaque assay on Vero E6 cells. Briefly, Vero E6 cells were seeded at a density of 1.3 × 105/ml in 12-well plates. Next day, cells were infected for 2 h with 10-fold serial dilutions of samples following which, cells were overlaid with 1 : 1 mixture of 2% agarose (Lonza) and 2X MEM medium (Gibco). On day 3, plaques were fixed using 10%

formaldehyde (Alfa Aesar) and stained using crystal violet (Sigma-Aldrich). Titers were calculated and reported as plaque-forming units (PFU) per ml.

Antiviral assays

For MHV infections, LR7 cells were seeded at a density of 1×10^5 cells per well in 12-well plates with a 10 mm diameter coverslip in it in DMEM containing 10% FBS, before to remove the medium and infect cells with MHV in DMEM at MOI 1. The inoculum was removed after 1 h, cells washed twice with 1x PBS (137 mM NaCl, 10 mM phosphate pH 7.4, 2.7 mM KCl) and fresh DMEM medium containing 2% FBS added. The tested drugs were added after another 1 h in the new medium and the cells were collected at 8 h p.i. Specifically, the cover slips were removed and fixed with 4% paraformaldehyde. After permeabilization using 0.2% Triton X-100 and subsequent blocking with PBS buffer containing 1% fetal calf serum, viral non-structural protein (nsp)2 and nsp3 were detected using the anti-nsp2/nsp3 antiserum, a kind gift from Susan Baker [26], followed by incubation with secondary antibody conjugated to Alexa-488 (Life Technologies). Fluorescence signals were captured with a Leica sp8 confocal microscope (Leica) and the nsp2/nsp3-positive cells were counted as infected cells. The cells in the 12 wells plates, in contrast, were harvested in 100 µl of 2× sample buffer (65.8 mM Tris–HCl, pH 6.8, 26.3% glycerol, 2.1% SDS and 0.01% bromophenol blue) for 30 min on ice, sonicated for 1 min and boiled. Equal protein amounts were separated by SDS-PAGE and after western blot, proteins were detected using specific antibodies against MHV N protein (a kind gift from Stuart Siddell, University of Bristol, UK [27]) and β-actin (Merck Millipore), and with secondary antibody conjugated to Alexa-488 (Life Technologies). Fluorescence signals were analyzed Odyssey imaging system (LI-COR) and signal intensities were normalized and quantified using the ImageJ software (NIH).

For SARS-CoV-2 infections, Vero E6 cells were seeded at a density of 6×10^4/well in 24 well plates in DMEM containing 10% FBS. Next day, plates were transferred to a Biosafety level 3 (BSL3) facility and replaced with 200 µl of DMEM medium containing 2% FBS and the virus inoculum (MOI 1). Following 2 h adsorption at 37 °C, virus inoculum was removed, after which cells were washed twice and replaced with DMEM media containing 10% FBS in combination with increasing concentrations of compounds or the equivalent volumes of DMSO as the control. Supernatants were harvested at 8 h p.i. and titrated using plaque assay and data were normalized to the non-treated control. In contrast, cells were washed with ice-cold PBS and lysed with TRIzol

reagent (Sigma) according to the manufacturer's instructions. First-strand cDNA was synthesized by using Moloney murine leukemia virus (M-MLV) reverse transcriptase and oligo (dT) (both from Invitrogen). Real-time PCR was performed using an CFX connect Thermocycler (Bio-Rad). The expression levels of SARS-CoV-2 nsp3 gene (forward primer: 5′-GCCTATACAGTTGAACTCGGT; reverse primer: 5′-CAATGCCCAGTGGTGTAAGT) were normalized to that of GAPDH (forward primer: 5′-AGCCACATCGCTCAGACAC; reverse primer: 5′-GCCCAATACGACCAAATCC) according to the comparative cycle threshold method used for quantification as recommended by the manufacture's protocol.

Statistical analyses

Statistical significances were evaluated using the two-tailed heteroscedastic t-test before calculating the p-values. Individual data points from each independent experiment were used for the calculation of the significance.

Results and discussion

Molecular docking of α-ketoamides onto 3CLpro

Based on the understanding of the molecular mode-of-action of drugs on their receptors, we initially computationally investigated approved drugs for their potential covalent interaction with the proteases of SARS-CoV-2, as a repurposing approach [28]. Drug families that attracted our attention were nitrile-containing gliptins and α-ketoamide bearing HCV NS3–4A protease inhibitors. Approved HCV NS3–4A protease inhibitors work through different mechanisms. N-Acylsulfonamides, e.g., danoprevir, inhibit the protease through a non-covalent mechanism, whereas α-ketoamides, e.g., boceprevir, forms a covalent bond with the active site serine. Therefore, only α-ketoamide HCV NS3–4A protease inhibitors were considered in our study as they interact covalently with the SARS-CoV-2 3CLpro. Computational docking of several HCV inhibitors, i.e., boceprevir, telaprevir and narlaprevir, into SARS-CoV-2 3CLpro indeed showed promising electrostatic and shape complementarity for the formation of a covalent bond between the active site Cys145 and the α-ketoamide group of these molecules and an overall complementary fit into the enlarged binding pocket. Amongst the three docked HCV NS3–4A protease inhibitors, boceprevir showed the highest ligand efficiency and was predicted to be superior to telaprevir and narlaprevir (Table S2). However, we could not further examine narlaprevir further because we did not have access to this molecule. None of the HCV NS3–4A

protease inhibitors, however, docked in the second SARS-CoV-2 protease PLpro with appreciable affinity.

Figure 2 shows the best docking poses of boceprevir and telaprevir. Interestingly, our docking results showed that the covalent bond and the α-ketoamide moiety could be stabilized by the formation of hydrogen bonds with the His41 and Gly143. These results encouraged us to perform further studies to assess the use of boceprevir and telaprevir as potential inhibitors of SARS-CoV-2 3CLpro.

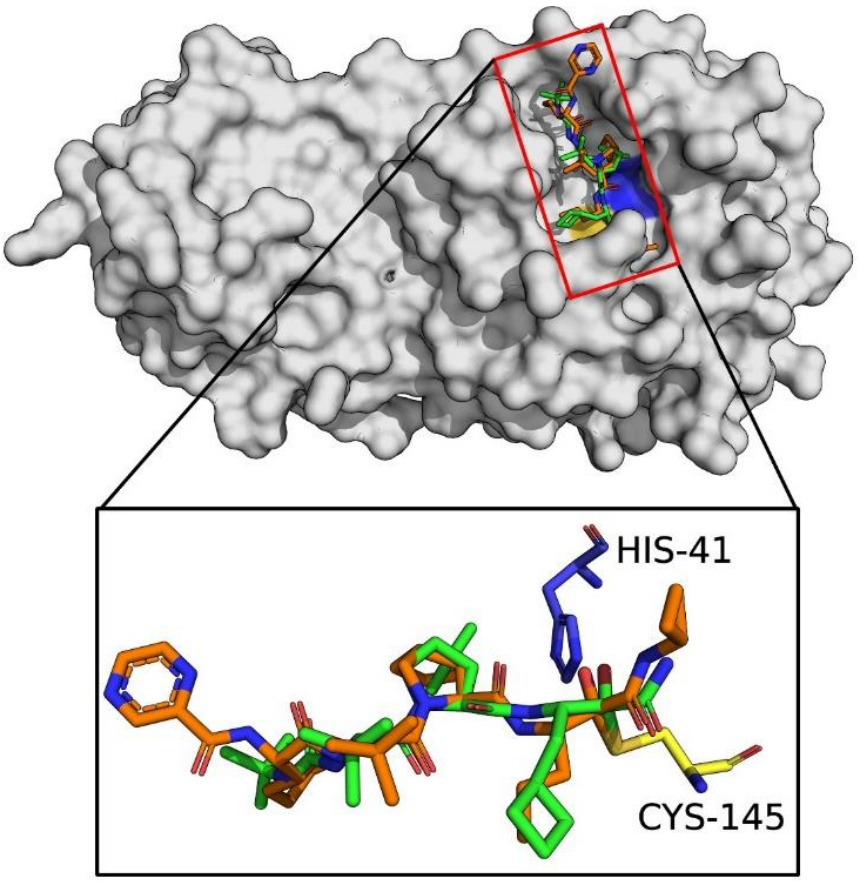


Figure 2. Docking of Boceprevir and Telaprevir onto SARS-CoV-2 3CLpro. The binding site of SARS-CoV-2 3CLpro (gray surface) with computationally docked Boceprevir (green sticks) and Telaprevir (orange sticks). The amino acids of the catalytic dyad are highlighted: Cys145 in yellow and His41 in blue sticks.

In vitro 3CLpro enzymatic inhibition by boceprevir and telaprevir

Encouraged by the promising docking results, we determined whether boceprevir and telaprevir could inhibit the enzymatic activity of 3CLpro. For this, we took advantage of an established fluorescence resonance energy transfer (FRET) assay [24]. SARS-CoV-2 3CLpro was

recombinantly expressed and purified, and using the FRET substrate 2-aminobenzoyl-SVTLQSG-Tyr(NO2)-R, the cleavage of this reporter peptide by 3CLpro was monitored. The IC50 of each compound was determined as described in Methods.

Boceprevir showed promising inhibition of 3CLpro with an IC50 of 1.59 μM (Fig. 3A). Telaprevir, in contrast, showed weak inhibition of cleavage activity, reaching 100% inhibition at 200 μM with an IC50 of 55.72 μM (Fig. 3B). These results, in agreement with the *in silico* predictions, indicate that boceprevir is a better inhibitor of 3CLpro.

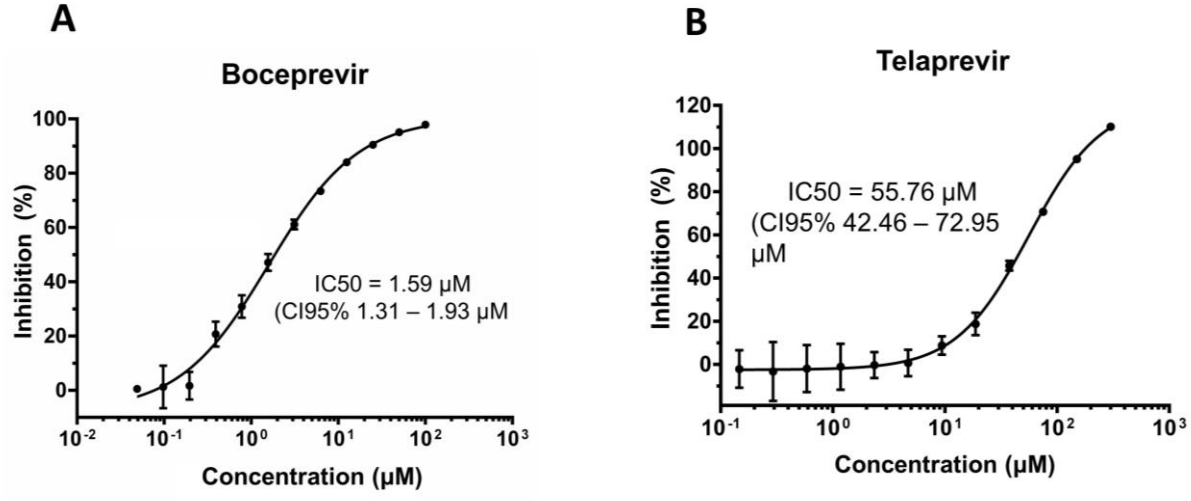


Figure 3. Boceprevir and Teleprevir are 3CLpro inhibitors. Inhibition of cleavage activity of SARS-CoV-2 3CLpro in the presence of increasing concentrations of A) Boceprevir and B) Telaprevir.

Crystal structures of SARS-CoV-2 3CLpro with boceprevir and telaprevir

In order to better understand how boceprevir and telaprevir inhibit the SARS-CoV-2 3CLpro and why their inhibitory activities differ, we determined the co-crystal structures of these drugs with 3CLpro. The crystal structure of SARS-CoV-2 3CLpro in complex with boceprevir was solved to 2.1 Å in space group C2 (Fig. S2 -PDB ID 6ZRU-). The asymmetric unit consists of a single monomer of SARS-CoV-2 3CLpro, but the active dimer is formed by a second molecule of SARS-CoV-2 3CLpro, which is related by crystallographic symmetry (Fig. S2B). Each monomer consists of three domains, domain I (residues 8–99), domain II (residues 100–183) and domain III (201–303), with the substrate-binding site, including the Cys145–His41 catalytic dyad, being located in a cleft between domains I and II.31,32 The substrate-binding site is made up of four conserved subsites: S1′, S1, S2 and S4.

The electron density map clearly shows boceprevir bound in the 3CLpro active site (Fig. S2A‡) where the carbonyl of the electrophilic α-ketoamide forms a 1.8 Å covalent bond with the sulphur of the catalytic Cys145, forming a S,O-acetale. The oxygen of the α-ketoamide forms important hydrogen bonds with the main chain amides of Cys145 and Gly143, occupying the oxyanion hole, and the hydroxyl group, resulting from the covalent addition to the α-ketoamide, forms a hydrogen bond with the sidechain of His41, which all stabilize the conformation. The cyclobutylmethyl group of boceprevir is positioned shallowly into the S1 pocket, angling up and away. The main chains of His164 and Glu166 form hydrogen bonds with the amide bonds on the main chain of boceprevir. The dimethyl-3-aza bicycle moiety inserts deeply into the S2 pocket, making extensive hydrophobic contacts with His41, Met49, Met165, Asp187, Arg188 and Gln189 (Fig. 4A and B).

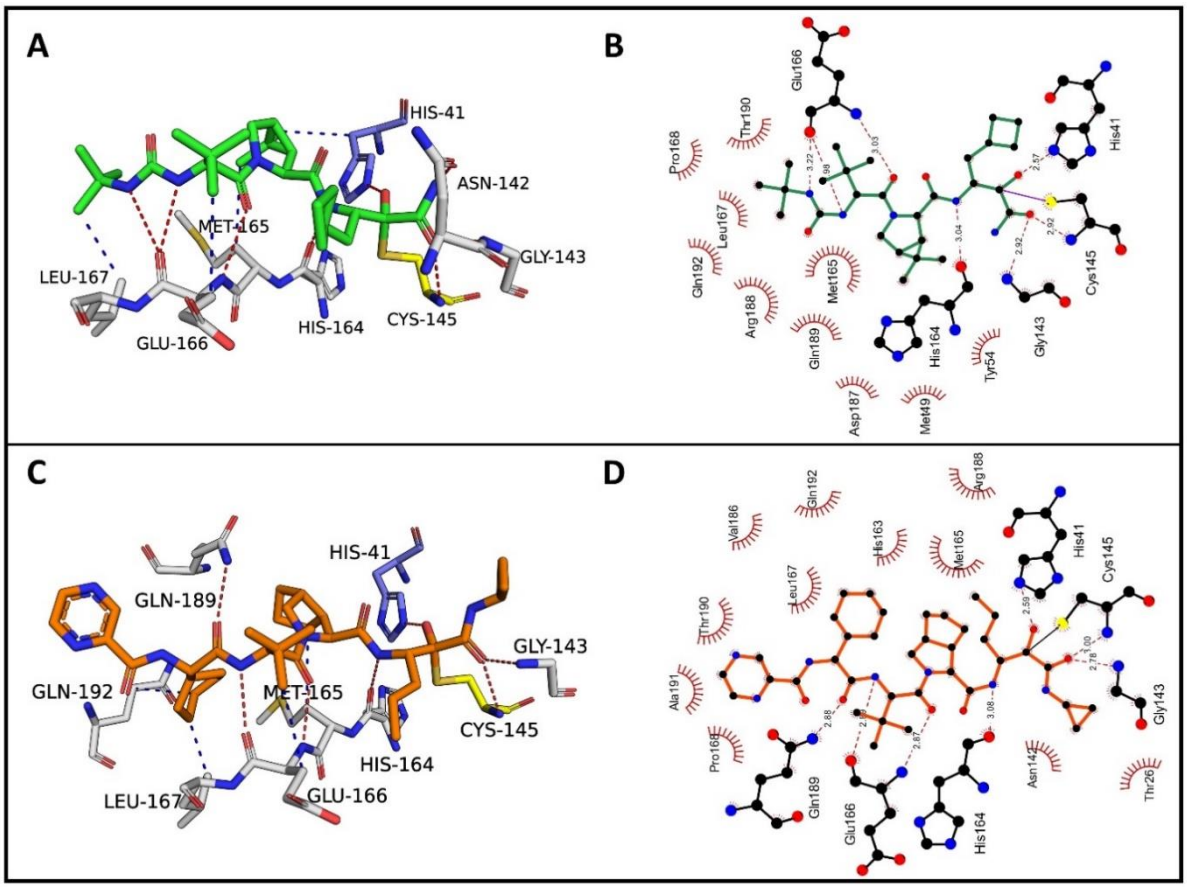


Figure 4. Boceprevir and Telaprevir binding to SARS-CoV-2 3CLpro. A) The Boceprevir (green sticks) and C) Telaprevir (orange sticks) binding sites showing the interactions with the key residues (white sticks). Hydrogen bonds are shown as blue lines, hydrophobic interactions as gray dashed lines. B) Schematic diagram of Boceprevir–3CLpro and D) Telaprevir-3CLpro interactions, made using Ligplot [29]. Green dashed lines represent hydrogen bonds, red curved lines indicate hydrophobic interactions.

The tert-butyl group is relatively solvent exposed, displaying only minor hydrophobic interactions with the sidechain of Glu166. Finally, the tert-butyl urea group orients deep into the S4 pocket, with the urea group being stabilized by several hydrogen bonds with the main chain oxygen of Glu166 and the tert-butyl undergoing hydrophobic interactions with the sidechains of Met165, Gln192, Leu167 and Pro168.

We compared the best pose from the virtual docking of boceprevir with our co-crystal structure PDB: 6ZRU, resulting in a root mean square deviation (RMSD) over all atoms of 1.47 Å, exhibiting almost perfect matching of the α-ketoamide warhead and sulfur atom comprising the covalent bond. The crystal structure of SARS-CoV-2 3CLpro in complex with telaprevir was solved at 2.1 Å resolution in space group C2 (Fig. S3 -PDB ID 6ZRT-). The electron density map clearly shows telaprevir bound in the 3CLpro active site (Fig. S3A) and it is very similar to the boceprevir–3CLpro complex (Fig. S2A). The α-ketoamide from telaprevir forms a covalent adduct in the same orientation as boceprevir and is stabilized by the same hydrogen bonds with Cys145, Gly143 and His41. However, telaprevir contains a cyclopropyl substituent on the ketoamide nitrogen, potentially providing steric and conformational hindrance for the α-ketoamide to orient itself in the S1′ pocket, making covalent adduct formation less effective. The propyl moiety of telaprevir protrudes deeper into the S1 site compared to the cyclobutylmethyl group of boceprevir, displacing an ordered water molecule present in the S1 site of the 3CLpro–boceprevir complex that is absent in the telaprevir complex (Fig. S2A and S3A). Similarly, to boceprevir, the main chain amides of telaprevir form hydrogen bonds with the backbone atoms of His614, Glu166 and Gln189.

The bicyclic moiety of telaprevir orients into the S2 pocket, making hydrophobic contacts with Arg188, Gln192 and Met165 (Fig. 4C and D). However, the penetration into the pocket is not as deep as the dimethyl-3-aza bicycle of boceprevir and the hydrophobic contacts are not as extensive. The tert-butyl group of telaprevir takes the same orientation as the corresponding group of boceprevir and is relatively solvent exposed, making only minor hydrophobic contacts with the sidechain of Glu166. The last difference is at the S4 pocket, where telaprevir has a cyclohexyl group, displaying hydrophobic interactions with Met165, Leu167 and Gln192, oriented deeply into the S4 pocket, followed by a relatively solvent exposed pyrazinamide moiety extending out of the active site, making only minor van der Waals interactions with Pro168, Thr190 and Ala191 (Fig. 4C and D). In contrast, boceprevir has the tert-butyl urea group that is fully located in the binding site.

We compared our virtual docking against our co-crystal structure PDB: 6ZRT. The best docked pose of telaprevir exhibited a RMSD value of 1.22 Å over all atoms against the crystal pose. The docking showed the highest deviation on the 2-pyrazine carboxamide moiety, indicating flexibility of this functional group when binding to 3CLpro, in line with the relatively high solvent exposure observed in the crystal structure. On the other hand, the orientation of the α-ketoamide warhead of the docked pose was nearly identical to the crystal pose(s), including the stabilizing hydrogen bonds with Cys145, His41, His164 and Gly143 (Fig. 4C and D).

As of now, apart from our released crystal-structures, other published and not published 3CLpro structures containing boceprevir [30–32] and telaprevir [31,33,34] are available. The other available boceprevir–3CLpro complex structures show highly similar binding poses to ours (Fig. S4A). Interestingly, the high-resolution structures (6WNP, 7K40, 7C6S and 7BRP) show a well-ordered water molecule coordinating an interaction between the amide group of the α-ketoamide and the backbone oxygen of Thr26 through hydrogen bonds. This water molecule is not present in the released 3CLpro – telaprevir structures, with the cyclopropyl group likely displacing it (Fig. S4B). Additionally, an overlay of available telaprevir structures shows conformational flexibility of the propyl moiety that is located in the S1 pocket, seemingly displacing the ordered water molecule in some structures (6ZRT, 7K6D) but not in others (7K6E, 7C7P and 6XQS). The pyrazinamide moiety of telaprevir also shows multiple conformations among the crystal structures.

Boceprevir and telaprevir in beta-CoV infections

Next, we tested whether boceprevir and telaprevir inhibited beta-CoV replication in cells. To this aim, we first perform a cytotoxic assay in the Vero E6 and LR7 cell lines to determine the maximal non-toxic concentration of these compounds that could be used in cells. Using the MTT method, we found out that the maximal non-toxic concentration of these drugs in these cell lines was 72.5 μM for both (Fig. 5A). To minimize the chance of possible side effects we decided to use a concentration of 40 μM of boceprevir and telaprevir for further cell studies.

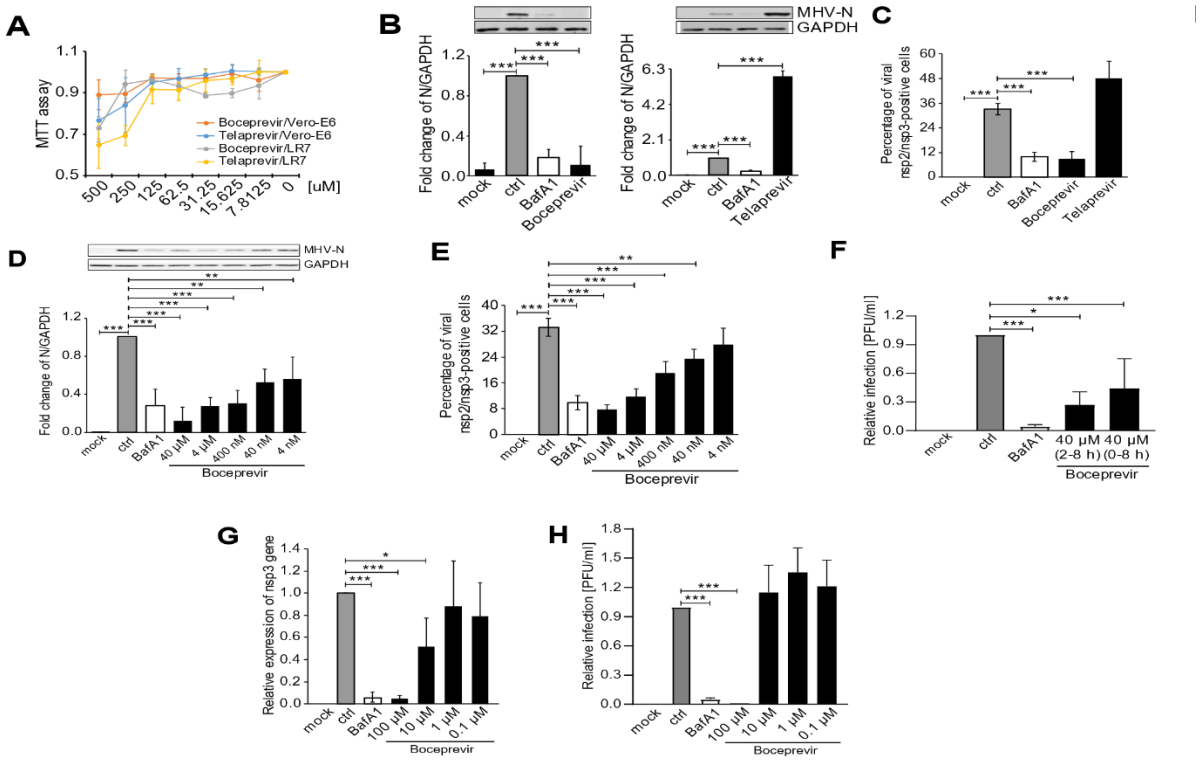


Figure 5. Boceprevir and Telaprevir in beta-CoV infections. A) The cytotoxicity of Boceprevir and Telaprevir in LR7 and Vero E6 cell lines. Cells were treated with the indicated doses of Boceprevir and Telaprevir for 6 h. Cell viability was subsequently measured using the MTT assay. All the MTT values were normalized to the 0.1% DMSO-only treatment, which represents 100% cell viability. B) LR7 cells were infected with MHV at MOI 1 for 2 h before adding 40 µM Boceprevir, 40 µM Telaprevir or 0.1% DMSO (ctrl) for another 6 h. Controls were cells not exposed to MHV (mock) or incubated with both MHV and 400 nM bafilomycin A1 (BafA1), which blocks virus cell entry. Proteins were separated by SDS-PAGE and western blot membranes probed with an antibody against either MHV N protein or GAPDH (top part). The N protein expression in each sample was quantified and normalized to the GAPDH signal. Results are expressed relative to the ctrl (low part). C) Cells treated as in panel B were processed for immunofluorescence using antibodies against MHV nsp2 and nsp3 proteins, and DAPI staining. The number of infected cells was subsequently determined. D) LR7 cells were infected with MHV at MOI 1 for 2 h before adding the indicated concentrations of Boceprevir for another 6 h. Cells treated with DMSO only (ctrl), BafA1-treated cells and cells not inoculated with MHV (mock) were used as controls. N protein levels were then examined as in panel B. E) Cells were treated as in panel D before statistically evaluating the percentage of infected cells by immunofluorescence as in panel C. F) Vero E6 cells were inoculated with SARS-CoV-2 at MOI 1, and treated with 40 µM Boceprevir at the same time or after 2h. At 8 h p.i., cell supernatants were collected and the number of produced infectious viral particles was determined using a plaque assay. Cells treated with DMSO only (ctrl), BafA1-treated cells and cells not inoculated with SARS-CoV-2 (mock) were used as controls. Results are expressed relative to the ctrl. G) Vero E6 cells were infected with SARS-CoV-2 at MOI 1 for 2 h before adding the indicated concentrations of Boceprevir for another 6h. Cells were lyses and the replication of SARS-CoV-2 was measured by assessing the expression levels of the mRNA encoding for nsp3 by RT-PCR and normalizing to those encoding for GAPDH. Cells treated with DMSO only (ctrl), BafA1-treated cells and cells not inoculated with SARS-CoV-2 (mock) were used as controls. Results are expressed relative to the ctrl. H) Culture supernatants of the samples analyzed in panel G were examined by plaque assay as in panel F. Results are

expressed relative to the ctrl. All data are represented as mean ± standard deviation of at least three independent experiments. Student T test was used to evaluate statistical differences and a p value ≤ 0.05 was considered significant with *p ≤ 0.05, **p ≤ 0.01 and ***p ≤ 0.001.

As a first analysis, we explored the antiviral effect of the two compounds in LR7 cells infected with mouse hepatitis virus (MHV), a model beta-CoV. We opted to add boceprevir and telaprevir 2 h p.i. The reason behind this choice was twofold. First, a specific inhibitor of 3CLpro should be able to block the viral replication, thus after virus cell entry has occurred. Second, a positive outcome of this approach would indicate that the compound can also be used to treat infected cells, something relevant from a therapeutic point of view. Thus, LR7 cells were infected with MHV at MOI 1 as described in the Methods section, and 40 μM boceprevir or 40 μM telaprevir were added 2 h p.i. The incubation was continued for an extra 6 h before processing the cells for either western blot (WB) with anti-MHV N protein antibodies, to assess viral protein production, or immunofluorescence (IF) with anti-nsp2/nsp3 antibodies, to determine the number of infected cells. As a positive control, we used bafilomycin A1, an inhibitor of the lysosomal H+-ATPase that increases the pH in the compartments of the endolysosomal system, blocking the cell entry of multiple viruses including CoV. This drug was therefore added at the same time as the virus inoculum and as expected, showed a strong reduction of MHV replication in both WB and IF readouts (Fig. 5B and C). The same assays also revealed that 40 μM telaprevir has no antiviral effect, but rather promoted the MHV infection (Fig. 5B and C). In contrast and seemingly in line with the results of the in vitro 3CLpro enzymatic assay, treatment with 40 μM boceprevir showed a significant inhibition of both MHV N protein production and the percentage of infected cells, to an extent similar to bafilomycin A1 treatment.

Next, we investigated whether the antiviral effect of boceprevir is dose-dependent, which would confirm its specificity. As shown in Fig. 5D and E, LR7 cells infected with MHV were treated with boceprevir at concentrations from 4 nM to 40 μM, with a serial dilution factor of 10. Interestingly, MHV N protein production was decreased by boceprevir in a dose-dependent manner, with a reduction of approximately 90% at a concentration of 40 μM and still of 40% at 4 nM, in comparison to the DMSO-treated infected cells (Fig. 5D). The dose-dependent inhibition of MHV replication by boceprevir was confirmed by determining the percentage of infected cells. This number was reduced by approximately 75% in cells subjected to 40 μM boceprevir and still by approximately 25% in those incubated with 40 nM, in comparison to the mock-treated infected cells (Fig. 5E). Altogether, the results show that boceprevir can efficiently inhibit MHV infection.

Then, we turned to SARS-CoV-2. Vero E6 cells were infected with SARS-CoV-2 at MOI and treated with 40 μM boceprevir at 2 h p.i. At 8 h p.i., cell culture supernatants were collected to titrate the progeny virus by plaque assay. We also added 40 μM boceprevir at the time of virus infection, to explore whether this could further enhance the observed effects of this compound. As shown in Fig. 5F and G, 40 μM boceprevir added at 2 h p.i. reduced the virus egression by approximately 50%, while its addition at the same time as the inoculum decreased virus progeny of approximately 75%.

While preparing our manuscript, an article appeared also showing that boceprevir can inhibit SARS-CoV-2 infection, and in this study the antiviral effects were started to be observed at a concentration of 1 μM. Thus, we infected Vero E6 cells with SARS-CoV-2 and at 2 h p.i., added boceprevir at concentrations ranging from 0.1 to 100 μM, with a serial dilution factor of 10. At 8 h p.i., cell culture supernatants were collected for the plaque assay while cells were lysed to extract the RNA and quantify viral replication by real-time PCR. This latter assay showed that 100 μM boceprevir effectively inhibited viral gene expression while 10 μM showed a decrease of approximately 50% (Fig. 5F and G). A lower concentration of boceprevir had no antiviral effect. Using the plaque assay, we only could observe a significant inhibition of SARS-CoV-2 egression only with the 100 μM boceprevir (Fig. 5H). Taken altogether, our results show that boceprevir has an evident antiviral effect against beta-CoV at concentrations ranging from 40–100 μM, but the working concentrations might be varying between different beta-CoV.

Discussion

We described the structural basis of the interaction of the HCV NS3–4A protease inhibitors boceprevir and telaprevir with the SARS-CoV-2 3CLpro and established that α-ketoamide HCV NS3–4A protease inhibitors can inhibit SARS-CoV-2 3CLpro. This is surprising since the two enzymes are evolutionary unrelated. The HCV NS3–4A is a serine protease depending on a catalytic triade, while SARS-CoV-2 is a cysteine protease with a catalytic dyad. Moreover, the substrate specificity of the two proteases is different and thus the shape and the electrostatic of the substrate pockets. We showed the added value of using computational docking in the identification of these drugs that could potentially be repurposed, displaying highly similar docking poses to the crystal structures. We showed that the HCV NS3–4A protease inhibitors boceprevir (and to a lesser extent telaprevir) are inhibitors of the SARS-CoV-2 3CLpro *in vitro*, with boceprevir also showing potent inhibition of viral replication *in vivo*. As seen in the co-crystal

structures, there are differences in subsite binding that could explain the increased inhibitory activity of boceprevir compared to telaprevir. Though the mode of inhibition of these two compounds with respect to the covalent adduct formation with the catalytic Cys145 is the same, the divergence mainly lies in the interactions with the S1′, S2 and S4 subpockets (Fig. 4, 6 and S4). Since covalent complex formation is dependent on the formation of an initial non-covalent precomplex, these distinctions could explain the difference in inhibitory potency. Boceprevir binds deeper into the S2 pocket and makes more extensive hydrophobic interactions than telaprevir. Additionally, in recently released, higher resolution structures of boceprevir, a well-ordered water molecule is shown to coordinate an interaction with the amide group of the α-ketoamide and the backbone oxygen of Thr26 through hydrogen bonds [32,35,36] (Fig. S4A), potentially stabilizing the covalent conformation. This interaction is absent in the telaprevir structures due to the cyclopropyl moiety occupying that space (Fig. S4B). Finally, telaprevir shows higher conformational flexibility in the S1 pocket and in the location of its pyrazinamide group. The less favorable interactions, coupled with higher degrees of conformational flexibility could explain the lower inhibitory potency of telaprevir in comparison to boceprevir.

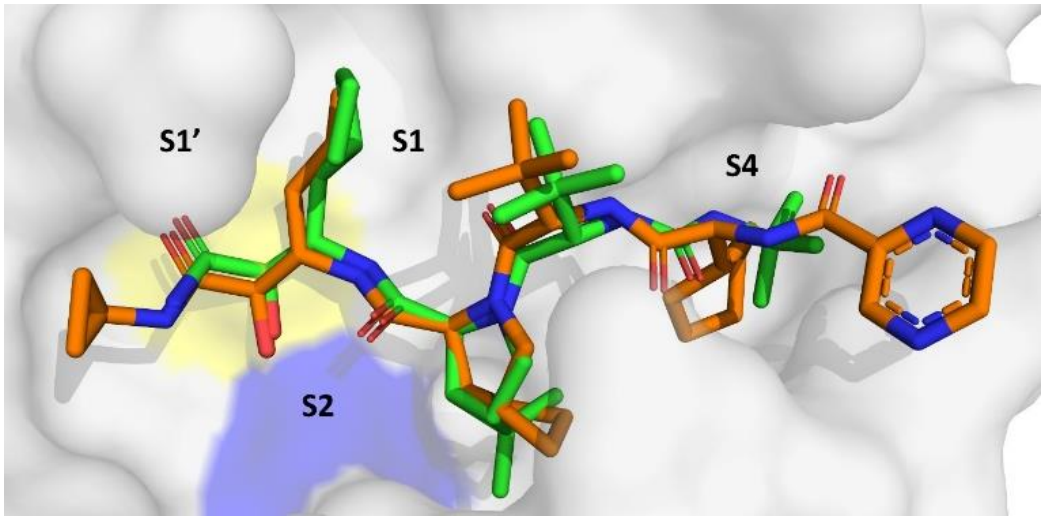


Figure 6. Alignments of Binding Modes of Boceprevir and Telaprevir. The SARS-CoV-2 3CLpro (gray surface) in complex with Boceprevir (green sticks) and Telaprevir (orange sticks). The residues od the catalytic dyad as shows in yellow surface for Cys145, and blue surface for His41.

Another α-ketoamide HCV NS3–4A protease inhibitor, narlaprevir also showed promising docking results, however, it could not be tested *in vitro* or in cell culture since it was not accessible to us. While preparing this manuscript, another research group published the discovery of the HCV NS3–4A protease inhibitor boceprevir as a 3CLpro inhibitor of SARS-CoV-2 [28]. Although our

study confirmed in gross their findings, it also reached some different conclusions. In particular we describe the co-crystal structure of boceprevir and 3CLpro, leading to a sound basis of understanding of the molecular interaction of the drug with the receptor. Our measured enzyme inhibition and cellular viral replication are higher than the ones reported in the previous study. The differences might be explained by a different time point of addition of boceprevir, and other details such a culture medium, and virus stock. Additionally, we found that not only boceprevir but also the FDA-approved telaprevir from the same drug class of HCV NS3–4A protease inhibitors, inhibits the 3CLpro *in vitro*. We describe here for the first time the molecular basis of the 3CLpro inhibition by telaprevir. The anti-SARS-CoV-2 mode-of-action of both drugs rely on the presence of an α-ketoamide moiety. Thus, we propose that the approved α-ketoamide HCV NS3–4A protease inhibitors boceprevir, telaprevir and possibly also other members of this family (e.g., narlaprevir) are promising COVID-19 repurposing candidates. While the potency of the herein proposed repurposed drugs in an *in vivo* setting are unknown, they deserve further attention as potential treatments for COVID-19 patients. Boceprevir, in particular, is safe for humans and is on the market since 2011 for the treatment of HCV infections and showed a limited number of side effects. Moreover, pharmacokinetics and pharmacodynamics in humans are well known [37]. The human plasma exposure of boceprevir was measured as Cmax 1.72 µg/ml, whereas Cmin 0.08 µg/ml, which is in the similar range as the herein reported cellular viral replication inhibition of 40 µM (21 µg/ml) [38]. Another potential application of this drug is based on the good potency for MHV, showing that it may represent an effective pan-anti-CoV inhibitor. It is well established that CoV infect farm animals. Thus, these compounds could be relevant for the cattle industry and possibly for future CoV epidemics.

Supplementary information

Fig S1. Sequence SARS-CoV-2 3CLpro ordered from Eurofins

CCATGGCGGCCGTACTGCAATCAGGTTTTCGCAAAATGGCGTTTCCATCGGGAAAA
GTCGAAGGCTGCATGGTTCAGGTTACATGTGGGACAACCACGCTGAATGGCCTGT
GGTTGGATGATGTGGTGTATTGTCCTCGTCACGTTATCTGCACAAGCGAAGATATGC
TGAATCCGAACTATGAGGACTTGCTGATTCGGAAATCCAATCACAACTTTCTGGTGC
AAGCGGGTAACGTGCAGTTACGCGTAATCGGCCATTCGATGCAGAACTGTGTGCT
GAAACTGAAAGTGGACACCGCGAATCCCAAAACCCCGAAATACAAGTTCGTCCGT
ATTCAACCAGGGCAGACCTTTAGCGTCCTCGCATGCTATAACGGCAGTCCGAGTG
GTGTGTATCAGTGTGCGATGCGTCCGAACTTCACCATCAAAGGCTCCTTTCTGAAC
GGGTCGTGTGGTAGCGTAGGCTTCAACATCGACTACGATTGCGTTAGCTTTTGCTA
CATGCATCACATGGAATTGCCGACTGGTGTCCATGCCGGTACTGATCTGGAAGGCA
ACTTCTATGGTCCCTTTGTTGATCGTCAGACCGCCCAAGCAGCGGGTACCGATACC
ACCATTACCGTGAATGTGCTCGCTTGGTTATATGCGGCTGTGATCAATGGAGATCG
CTGGTTTCTGAATCGCTTCACGACCACGCTTAACGACTTCAATCTCGTCGCAATGAA
GTACAACTACGAACCTCTGACTCAGGATCATGTGGATATTCTGGGTCCGTTATCTGC
TCAGACGGGCATTGCCGTACTGGACATGTGCGCCTCACTGAAGGAGTTACTGCAG
AACGGGATGAATGGACGCACGATTTTGGGCTCTGCACTTCTTGAGGACGAATTCAC
TCCGTTTGATGTTGTCCGCCAATGCAGCGGCGTTACGTTTCAGCTCGAG

Table S1. Data collection and refinement statistics

Data collection	SARS-CoV-2 Mpro - Boceprevir (PDB ID: 6ZRU)	SARS-CoV-2 Mpro - Telaprevir (PDB ID: 6ZRT)
Space group	I 1 2 1	C 1 2 1
Cell dimensions		
a, b, c (Å)	113.76, 53.53, 45.88	109.68, 54.99, 47.93
α, β, γ (°)	90.00, 101.52, 90.00	90.00, 101.42, 90.00
Completeness (%)	100.0 (100.0)	99.3 (98.5)
Rmerge	0.040 (0.702)	0.050 (0.549)
<I/σ(I)>	23.9 (2.8)	22.1 (3.8)
Redundancy	6.8 (7.0)	6.4 (6.7)
Refinement		
Resolution	44.96 – 2.10 (2.16 – 2.10)	47.03 – 2.10 (2.16 – 2.10)
No. reflections	15173 (1131)	15497 (1130)
Rwork / Rfree	0.188 / 0.215	0.199 / 0.237
No. Atoms		
Protein	2340	2340
Ligand/Ion	45	53
Water	23	45
B-Factors (Å2)		
Protein	55.27	43.07
Ligand/Ion	61.67	48.25
Solvent	45.28	42.31
RMS deviations		
Bond lengths (Å)	0.009	0.009
Bond Angles (°)	1.654	1.594

Table S2. Best pose docking results of HCV NS3-4A protease inhibitors

Ligand	PLPChemscore	Ligand Efficiency
Boceprevir	125.94	3.31
Telaprevir	162.25	3.25
Narlaprevir	95.84	1.92

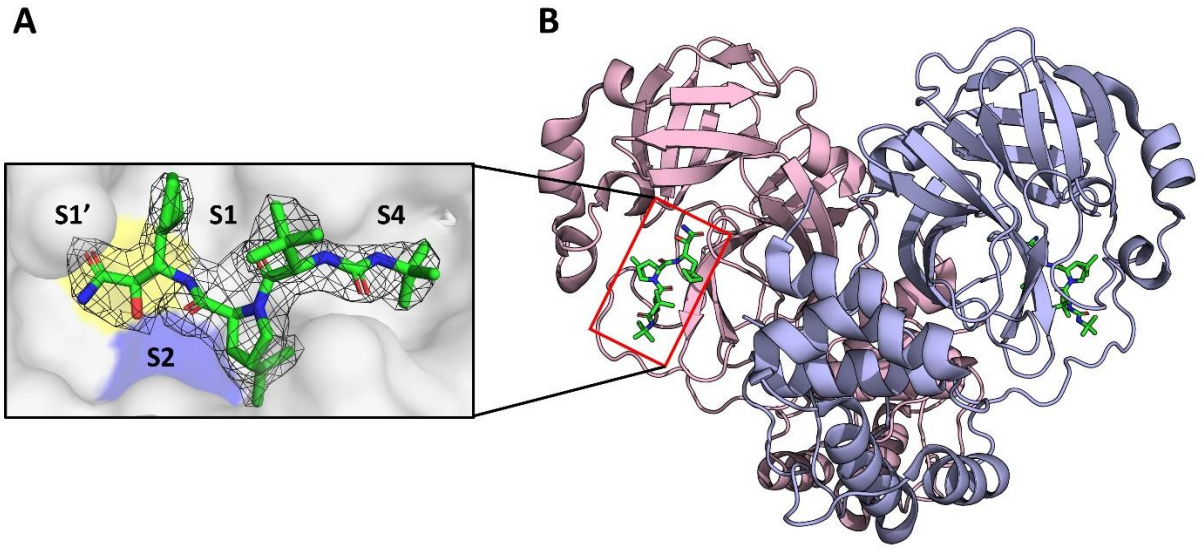


Figure S2. Electron density of Boceprevir. A) Magnified view of the substrate binding pocket (surface representation) with subsites S1', S1, S2 and S4 indicated. Boceprevir is shown as green sticks, enmeshed by the 2Fo-Fc map contoured at 1.0 σ. Cys145 in yellow surface and His41 in blue surface. B) Cartoon representation of the biological assembly (homodimer) of SARS-Cov-2 3CLpro covalently bound to Boceprevir presented as green sticks. Protomer A is pink, protomer B is purple.

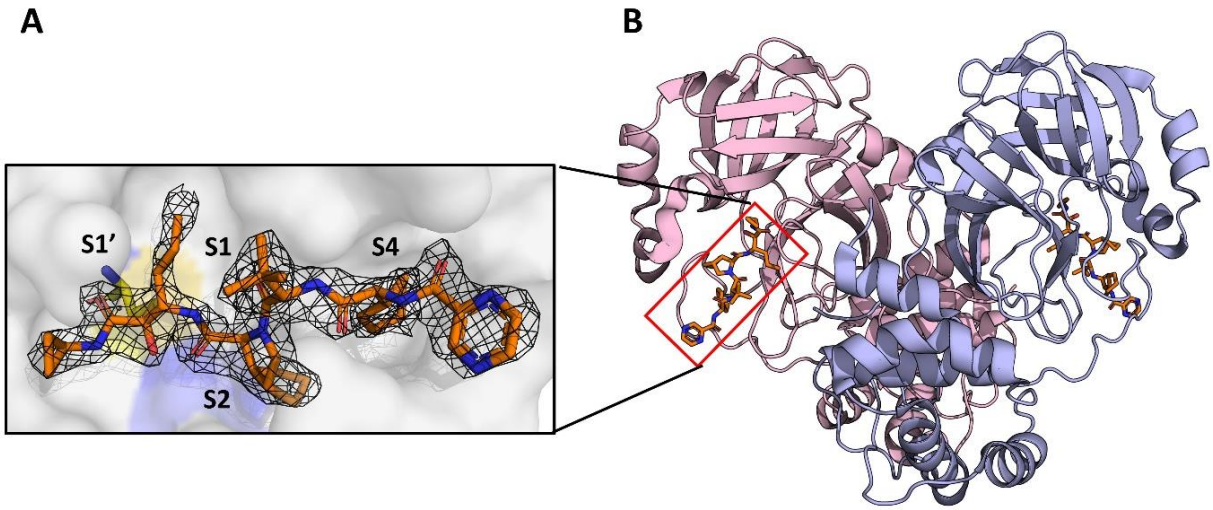


Figure S3. Electron density of Telaprevir. A) Magnified view of the substrate binding pocket (surface representation) with subsites S1', S1, S2 and S4 indicated. Telaprevir is shown as orange sticks, enmeshed by the 2Fo-Fc map contoured at 1.0 σ. Cys145 in yellow surface and His41 in blue surface. B) Cartoon representation of the biological assembly (homodimer) of SARS-Cov-2 3CLpro covalently bound to Telaprevir, presented as orange sticks. Protomer A is pink, protomer B is purple.

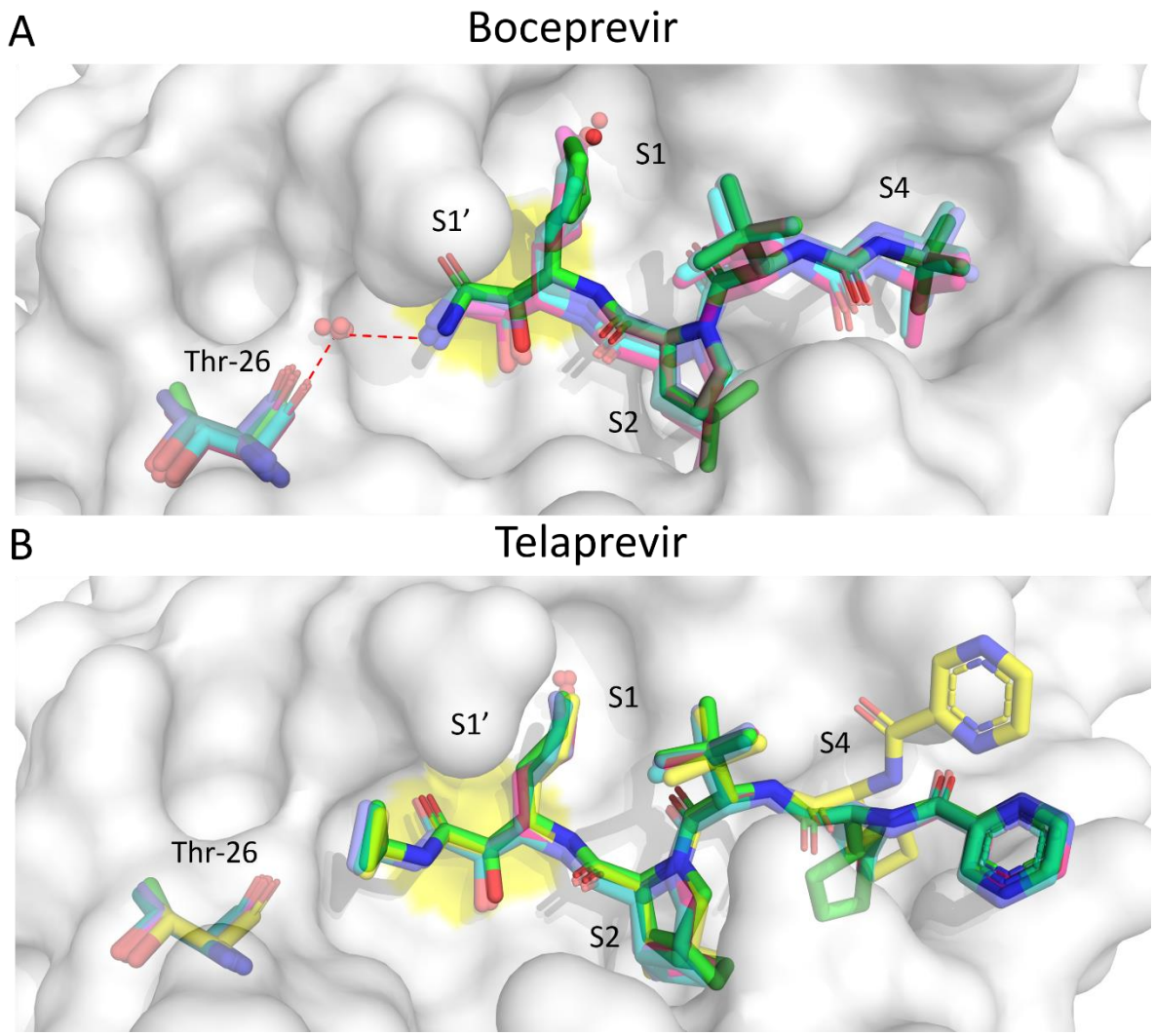


Figure S4. Structural alignment of available crystal poses of 3CLpro-Boceprevir and Telaprevir complexes. A) Overlay of the reported SARS-CoV-2 3CLpro (white surface) structure in complex with Boceprevir (6ZRU) and 5 available crystals (7K40, 7C6S, 6XQU,7BRP and 6WNP). The Boceprevir molecules are represented as sticks (6ZRU solid green sticks, 7K40 slate sticks, 7C6S cyan, 6XQU light magenta,7BRP deep teal and 6WNP warm pink), Cys145 is shown as a yellow surface and Thr26 as sticks. Waters are represented as spheres. The water mediated hydrogen bond between Thr26 and Boceprevir (6WNP) is indicated with dashed lines. B) Overlay of the reported SARS-CoV-2 3CLpro (white surface) structure in complex with Telaprevir (6ZRT) and 4 available crystals (7C7P, 7K6E, 7K6D and 6XQS). The Telaprevir molecules are represented as sticks (6ZRT solid green sticks, 7C7P yellow sticks, 7K6E slate, 7K6D hot pink and 6XQS teal), Cys145 is shown as a yellow surface and Thr26 as sticks. Waters are represented as spheres.

References

1. Chen N, Zhou M, Dong X, Qu J, Gong F, Han Y, et al. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. The Lancet. 2020;395: 507–513. doi:10.1016/S0140-6736(20)30211-7

2. Wang D, Hu B, Hu C, Zhu F, Liu X, Zhang J, et al. Clinical Characteristics of 138 Hospitalized Patients With 2019 Novel Coronavirus–Infected Pneumonia in Wuhan, China. JAMA. 2020;323: 1061. doi:10.1001/jama.2020.1585

3. Epidemiological and clinical features of the 2019 novel coronavirus outbreak in China | medRxiv. [cited 3 May 2020]. Available: https://www.medrxiv.org/content/10.1101/2020.02.10.20021675v2

4. Chakraborty I, Maity P. COVID-19 outbreak: Migration, effects on society, global environment and prevention. Sci Total Environ. 2020;728: 138882. doi:10.1016/j.scitotenv.2020.138882

5. Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. The Lancet. 2020;395: 497–506. doi:10.1016/S0140-6736(20)30183-5

6. ul Qamar MT, Alqahtani SM, Alamri MA, Chen L-L. Structural basis of SARS-CoV-2 3CLpro and anti-COVID-19 drug discovery from medicinal plants†. Journal of Pharmaceutical Analysis. 2020 [cited 3 May 2020]. doi:10.1016/j.jpha.2020.03.009

7. Zhou P, Yang X-L, Wang X-G, Hu B, Zhang L, Zhang W, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature. 2020;579: 270–273. doi:10.1038/s41586-020-2012-7

8. Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, et al. A Novel Coronavirus from Patients with Pneumonia in China, 2019. N Engl J Med. 2020;382: 727–733. doi:10.1056/NEJMoa2001017

9. Wu F, Zhao S, Yu B, Chen Y-M, Wang W, Song Z-G, et al. A new coronavirus associated with human respiratory disease in China. Nature. 2020;579: 265–269. doi:10.1038/s41586-020-2008-3

10. Perlman S, Netland J. Coronaviruses post-SARS: update on replication and pathogenesis. Nat Rev Microbiol. 2009;7: 439–450. doi:10.1038/nrmicro2147

11. Fehr AR, Perlman S. Coronaviruses: An Overview of Their Replication and Pathogenesis. In: Maier HJ, Bickerton E, Britton P, editors. Coronaviruses: Methods and Protocols. New York, NY: Springer; 2015. pp. 1–23. doi:10.1007/978-1-4939-2438-7_1

12. Anand K, Ziebuhr J, Wadhwani P, Mesters JR, Hilgenfeld R. Coronavirus Main Proteinase (3CLpro) Structure: Basis for Design of Anti-SARS Drugs. Science. 2003;300: 1763–1767. doi:10.1126/science.1085658

13. Mielech AM, Chen Y, Mesecar AD, Baker SC. Nidovirus papain-like proteases: multifunctional enzymes with protease, deubiquitinating and deISGylating activities. Virus Res. 2014;194: 184–190. doi:10.1016/j.virusres.2014.01.025

14. Shin D, Mukherjee R, Grewe D, Bojkova D, Baek K, Bhattacharya A, et al. Papain-like protease regulates SARS-CoV-2 viral spread and innate immunity. Nature. 2020;587. doi:10.1038/s41586-020-2601-5

15. Groves M, Domling A, Moreno AJR, Romero AR, Neochoritis C, Velasco-Velázquez M. Gliptin Repurposing for COVID-19. 2020 [cited 5 Jul 2020]. doi:10.26434/chemrxiv.12110760.v1

16. Tuley A, Fast W. The Taxonomy of Covalent Inhibitors. Biochemistry. 2018;57: 3326–3337. doi:10.1021/acs.biochem.8b00315

17. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, et al. UCSF Chimera--a visualization system for exploratory research and analysis. J Comput Chem. 2004;25: 1605–1612. doi:10.1002/jcc.20084

18. Jones G, Willett P, Glen RC, Leach AR, Taylor R. Development and validation of a genetic algorithm for flexible docking. J Mol Biol. 1997;267: 727–748. doi:10.1006/jmbi.1996.0897

19. Dessau MA, Modis Y. Protein Crystallization for X-ray Crystallography. JoVE (Journal of Visualized Experiments). 2011; e2285. doi:10.3791/2285

20. Evans PR, Murshudov GN. How good are my data and what is the resolution? Acta Crystallogr D Biol Crystallogr. 2013;69: 1204–1214. doi:10.1107/S0907444913000061

21. Winn MD, Ballard CC, Cowtan KD, Dodson EJ, Emsley P, Evans PR, et al. Overview of the CCP4 suite and current developments. Acta Cryst D. 2011;67: 235–242. doi:10.1107/S0907444910045749

22. Emsley P, Cowtan K. Coot: model-building tools for molecular graphics. Acta Crystallogr D Biol Crystallogr. 2004;60: 2126–2132. doi:10.1107/S0907444904019158

23. Murshudov GN, Skubák P, Lebedev AA, Pannu NS, Steiner RA, Nicholls RA, et al. REFMAC5 for the refinement of macromolecular crystal structures. Acta Cryst D. 2011;67: 355–367. doi:10.1107/S0907444911001314

24. Blanchard JE, Elowe NH, Huitema C, Fortin PD, Cechetto JD, Eltis LD, et al. High-Throughput Screening Identifies Inhibitors of the SARS Coronavirus Main Proteinase. Chemistry & Biology. 2004;11: 1445–1453. doi:10.1016/j.chembiol.2004.08.011

25. Biacchesi S, Pham QN, Skiadopoulos MH, Murphy BR, Collins PL, Buchholz UJ. Infection of Nonhuman Primates with Recombinant Human Metapneumovirus Lacking the SH, G, or M2-2 Protein Categorizes Each as a Nonessential Accessory Protein and Identifies Vaccine Candidates. J Virol. 2005;79: 12608–12613. doi:10.1128/JVI.79.19.12608-12613.2005

26. Schiller JJ, Kanjanahaluethai A, Baker SC. Processing of the Coronavirus MHV-JHM Polymerase Polyprotein: Identification of Precursors and Proteolytic Products Spanning 400 Kilodaltons of ORF1a. Virology. 1998;242: 288–302. doi:10.1006/viro.1997.9010

27. Schwarz B, Routledge E, Siddell SG. Murine coronavirus nonstructural protein ns2 is not essential for virus replication in transformed cells. J Virol. 1990;64: 4784–4791.

28. Dömling A, Gao L. Chemistry and Biology of SARS-CoV-2. Chem. 2020;6: 1283–1295. doi:10.1016/j.chempr.2020.04.023

29. Laskowski RA, Swindells MB. LigPlot+: multiple ligand-protein interaction diagrams for drug discovery. J Chem Inf Model. 2011;51: 2778–2786. doi:10.1021/ci200227u

30. Fu L, Ye F, Feng Y, Yu F, Wang Q, Wu Y, et al. Both Boceprevir and GC376 efficaciously inhibit SARS-CoV-2 by targeting its main protease. Nature Communications. 2020;11. doi:10.1038/s41467-020-18233-x

31. Kneller DW, Galanie S, Phillips G, O'Neill HM, Coates L, Kovalevsky A. Malleability of the SARS-CoV-2 3CL Mpro Active-Site Cavity Facilitates Binding of Clinical Antivirals. Structure. 2020;28: 1313-1320.e3. doi:10.1016/j.str.2020.10.007

32. Bank RPD. RCSB PDB - 7K40: Crystal Structure of SARS-CoV-2 Main Protease (3CLpro/Mpro) in Complex with Covalent Inhibitor Boceprevir at 1.35 A Resolution. [cited 3 Dec 2020]. Available: https://www.rcsb.org/structure/7K40

33. Bank RPD. RCSB PDB - 7K6D: SARS-CoV-2 Main Protease Co-Crystal Structure with Telaprevir Determined from Crystals Grown with 40 nL Acoustically Ejected Mpro Droplets at 1.48 A Resolution (Cryo-protected). [cited 3 Dec 2020]. Available: https://www.rcsb.org/structure/7K6D

34. Bank RPD. RCSB PDB - 7K6E: SARS-CoV-2 Main Protease Co-Crystal Structure with Telaprevir Determined from Crystals Grown with 40 nL Acoustically Ejected Mpro Droplets at 1.63 A Resolution (Direct Vitrification). [cited 3 Dec 2020]. Available: https://www.rcsb.org/structure/7K6E

35. Jin Z, Du X, Xu Y, Deng Y, Liu M, Zhao Y, et al. Structure of M pro from SARS-CoV-2 and discovery of its inhibitors. Nature. 2020;582: 289–293. doi:10.1038/s41586-020-2223-y

36. RCSB PDB - 6WNP: X-ray Structure of SARS-CoV-2 main protease bound to Boceprevir at 1.45 A. [cited 21 Aug 2020]. Available: https://www.rcsb.org/structure/6WNP

37. E V, P B, V S. Pharmacokinetics of new oral hepatitis C antiviral drugs. Expert Opin Drug Metab Toxicol. 2012;9: 5–16. doi:10.1517/17425255.2013.729577

38. Johnson M, Borland J, Chen S, Savina P, Wynne B, Piscitelli S. Effects of boceprevir and telaprevir on the pharmacokinetics of dolutegravir. Br J Clin Pharmacol. 2014;78: 1043–1049. doi:10.1111/bcp.12428

Summary of the thesis and perspectives

Drug discovery is a process that aims to identify drug candidates by a thorough evaluation of the biological activity of small synthetic molecules or biomolecules. The modern drug discovery process includes identifying the disease to be treated and its unmet medical need; selecting a druggable molecular target and validating it; developing *in vitro* assays followed by a high throughput screening of compound libraries to identify hits towards the target; and hit optimization to generate lead compounds with adequate potency and selectivity towards the biological target in vitro, and that demonstrate efficacy in animal models. Subsequently, the lead compounds are further optimized to improve their potency and pharmacokinetics before moving forward with the clinical development. Computational strategies are now necessary tools for speeding up the drug discovery process. The use of computational approaches during specific stages of the drug discovery process, ranging from earliest stages to the application of Artificial Intelligence (AI) is described in Chapter 1.

Moreover, with the increasing availability of computational power, there has been a huge improvement in the speed and reliability of techniques for protein modeling, chemical space exploration, and biological target selection and validation. As a result of the introduction of AI and Machine Learning (ML) algorithms, current and future efforts are focused on making more precise predictions about the biological effects of drug candidates. Calculations of binding affinity, the evaluation of protein target conformational changes, potential off-target binding effects and the bioavailability modeling are just a few examples of how computational approaches support drug discovery.

The use of *in silico* methodologies for the design of potential drugs for biological validation targets is demonstrated in Chapter 2. The chapter details a *in silico* workflow for the identification of potential high-affinity antagonists of CD44, ranging from a structural analysis of the target to the analysis of ligand-protein interactions by means of molecular dynamics (MD). CD44 is a target for the development of new anti-cancer therapies, since it promotes metastasis, chemoresistance and stemness in various types of cancer. A common N-terminal domain found on all CD44 isoforms binds to hyaluronic acid (HA). We have identified a subdomain that binds to 1,2,3,4-tetrahydroisoquinoline (THQ)-containing compounds, and that is adjacent to HA interaction residues by means of analyzing 30 crystal structures of the HA-binding domain (CD44HAbd). We have generated a new library of 168,190 molecules with the THQ motif by computational combinatorial chemistry (CCC), and we have compared their conformers to a pharmacophore containing the key features of the crystallographic THQ binding mode. About 0.01 percent of the

compounds matched the pharmacophore, and were studied through computational docking and MD. In explicit-solvent MD simulations, we have found two compounds, Can125 and Can159, that bound to human (h) CD44HAbd, and thus may elicit the CD44 blockage. These compounds can be easily synthesized for activity testing using multicomponent reactions (MCR), and the binding mode reported here could aid in the development of more potent CD44 antagonists.

In Chapter 3, we have tested the shape-guided algorithm on a dataset of 208 macrocycles that were carefully chosen based on their structural complexity. Macrocycles are used to target proteins that are otherwise difficult to target due to a lack of hydrophobic cavities and extended featureless surfaces. Increasing efforts by computational chemists have resulted in the development of effective software to overcome the torsional and conformational restrictions imposed by macrocyclization. Since 1986, drug designers and crystallographers in the Roche biostructural community have been constantly updating Moloc, which is an efficient algorithm with a focus on high interactivity. We have quantified the accuracy, diversity, speed, exhaustiveness, and sampling efficiency of the dataset in an automated fashion, and we have compared them with four commercial packages (Prime, MacroModel, molecular operating environment and MD) and four open-access packages (the experimental-torsion distance geometry, with additional "basic knowledge" alone and with the Merck molecular force field minimization or the universal force field minimization, the Cambridge Crystallographic Data Centre conformer generator, and the conformator). Moloc displayed the highest sampling efficiency and exhaustiveness without producing thousands of conformations or random ring splitting into two half-loops. Besides, Moloc displayed the ability to produce highly accurate globular or flat conformations in a similar fashion as Prime, MacroModel and MD for 75 % of the studied dataset of macrocyles. These findings will allow further *in silico* evaluations of macrocycles as potential bioactive compounds. Further, we have identified the characteristics that need to be improved for the development of new tools for macrocycle sampling and design, such as those based on machine learning in order to predict the macrocycle conformations.

In Chapter 4, we describe a detailed reverse docking protocol for the identification of potential targets for 4-hydroxycoumarin (4-HC). Molecular docking is a useful and powerful computational method for the identification of potential interactions between small molecules and pharmacological targets. The ability of one or a few compounds to bind a large dataset of proteins is assessed *in silico* in reverse docking. This strategy is useful for identifying molecular targets of orphan bioactive compounds, discovering alternative drug indications (repurposing), and predicting drug toxicity. Our findings have revealed that RAC1 is a target of 4-HC, which helps to explain some of the biological effects of 4-HC on cancer cells. The strategy described in this

chapter can easily be applied to other compounds and protein datasets overcoming bottlenecks in molecular docking protocols, particularly in reverse docking approaches.

Lastly, Chapter 5 demonstrates how computational methods and experimental results can be used to repurpose compounds that could be used as treatments for COVID-19. The HCV NS3–4A protease inhibitors boceprevir and telaprevir were discovered in this chapter as SARS-CoV-2 main protease (3CLpro) inhibitors. We focused on the repurposing of drugs against SARS-CoV-2 because COVID-19 cases are rapidly increasing, resulting in an increasing mortality, and paralyzing the global economy. In contrast to *de novo* drug discovery, which takes years to go from concept to pre-clinical to commercialization, drug repurposing may constitute a short-term solution. In order to select those drugs with high probability of being active against 3CLpro, we have performed a covalent docking analysis. The selected drugs were assessed in enzyme inhibition assays and co-crystalized with the target to corroborate the hypothesis that α-ketoamide drugs can covalently bind to the active site cysteine of the SARS-CoV-2 3CLpro. Finally, we have established that boceprevir, and not telaprevir, inhibits the replication of SARS-CoV-2 and the mouse hepatitis virus (MHV), another coronavirus, in cell culture. According to our findings, the HCV drug boceprevir should be tested clinically for COVID-19 or other coronaviral infections.

Summarizing, these chapters show the importance, application, and limitations of computational methods in the state-of- the-art drug design process.

Samenvatting van het proefschrift en perspectieven

Geneesmiddelenontdekking is een proces dat tot doel heeft kandidaat-geneesmiddelen te identificeren door een grondige evaluatie van de biologische activiteit van kleine synthetische moleculen of biomoleculen. Het moderne medicijnontdekkingsproces omvat het identificeren van de te behandelen ziekte en de onvervulde medische behoefte; het selecteren van een 'medicijnbaar' moleculair doelwit en het valideren ervan; het ontwikkelen van in-vitro-analyses gevolgd door een screening met een hoge verwerkingscapaciteit van verbindingsbibliotheken, om treffers naar het doelwit toe te identificeren; en hit-optimalisatie, om leidende verbindingen te genereren met voldoende potentie en selectiviteit gericht op het biologische in-vitrodoel, en die werkzaamheid aantonen in diermodellen. Vervolgens worden de leidende verbindingen verder geoptimaliseerd, om hun potentie en farmacokinetiek te verbeteren, alvorens verder te gaan met de klinische ontwikkeling. Computationele strategieën zijn nu noodzakelijke hulpmiddelen om het ontdekkingsproces van geneesmiddelen te versnellen. Het gebruik van computationele benaderingen tijdens specifieke fasen van het medicijnontdekkingsproces, reikend van de vroegste stadia tot de toepassing van kunstmatige intelligentie (*AI*), wordt beschreven in Hoofdstuk 1.

Daarnaast is er, met de toenemende beschikbaarheid van rekenkracht, een enorme verbetering gekomen in de snelheid en betrouwbaarheid van de technieken voor eiwitmodellering, verkenning van de chemische ruimte, en biologische doelselectie en -validatie. Ten gevolge van de invoering van algoritmen voor *AI*- en Machinaal Leren (ML), zijn de huidige en toekomstige inspanningen gericht op het maken van nauwkeurigere voorspellingen omtrent de biologische effecten van kandidaat-geneesmiddelen. De berekeningen van bindingsaffiniteit, de evaluatie van conformationele veranderingen van het eiwitdoelwit, van potentiële *off-target* bindingseffecten en de modellering van biologische beschikbaarheid zijn slechts enkele voorbeelden van hoe computationele benaderingen de ontdekking van geneesmiddelen bevorderen.

Het gebruik van in-silicomethodologieën voor het ontwerpen van potentiële geneesmiddelen voor biologische validatiedoelen wordt getoond in Hoofdstuk 2. Het hoofdstuk detailleert een in-silicowerkstroom voor de identificatie van potentiële CD44-antagonisten met een hoge affiniteit, reikend van een structurele analyse van het doelwit tot de analyse van ligand-eiwitinteracties door middel van moleculaire dynamica (MD). CD44 is een doelwit voor de ontwikkeling van nieuwe antikankertherapieën, omdat het metastase, chemoresistentie en 'stamachtigheid' bij verschillende soorten kanker in de hand werkt. Een gemeenschappelijk N-terminaal domein dat op alle CD44-isovormen wordt aangetroffen, bindt aan hyaluronzuur (HA). We hebben een

subdomein geïdentificeerd dat bindt aan verbindingen die 1,2,3,4-tetrahydro-isochinoline (THQ) bevatten, en dat grenst aan HA-interactieresiduen door middel van een analyse van 30 kristalstructuren van het HA-bindende domein (CD44HAbd). We hebben een nieuwe bibliotheek opgebouwd van 168.190 moleculen met het THQ-motief door middel van computationele combinatorische chemie (CCC), en we hebben hun conformers vergeleken met een farmacofoor die de belangrijkste kenmerken van de kristallografische THQ-bindingsmodus bevat. Ongeveer 0,01 procent van de verbindingen kwam overeen met de farmacofoor, en werd bestudeerd door middel van computationeel koppelen en MD. In MD-simulaties met expliciete oplosmiddelen hebben we twee verbindingen gevonden, Can125 en Can159, die aan menselijk (h) CD44HAbd bonden, en die dus de CD44-blokkering kunnen veroorzaken. Deze verbindingen kunnen gemakkelijk gesynthetiseerd worden voor het testen van activiteit met behulp van multicomponent-reacties (MCR), en de hier gerapporteerde bindingsmodus zou kunnen helpen bij de ontwikkeling van krachtigere CD44-antagonisten.

In Hoofdstuk 3 hebben we het vormgestuurde algoritme getest op een dataset van 208 macrocylische verbindingen die zorgvuldig gekozen werden op basis van hun structurele complexiteit. Macrocylische verbindingen worden gebruikt om eiwitten te 'targeten' die anders moeilijk te bereiken zijn vanwege een gebrek aan hydrofobe holtes en uitgestrekte, karakterloze oppervlakken. Toenemende inspanningen van computationele scheikundigen hebben tot de ontwikkeling geleid van effectieve software om de door de macrocyclisatie opgelegde torsie- en conformationele beperkingen te overwinnen. Sinds 1986 hebben geneesmiddelenontwerpers en kristallografen in de biostructurele Roche-gemeenschap van Moloc voortdurend bijgewerkt. Moloc is een efficiënt algoritme met een focus op hoge interactiviteit. We hebben de nauwkeurigheid, diversiteit, snelheid, volledigheid en bemonsteringsefficiëntie van de dataset op een geautomatiseerde wijze gekwantificeerd, en we hebben ze vergeleken met vier betalende pakketten (Prime, MacroModel, moleculaire werkomgeving en MD) en vier *open access*-pakketten (de experimentele torsie- en afstandsgeometrie, met alleen aanvullende "basiskennis" en met de minimalisatie van het moleculaire krachtveld van Merck of de minimalisatie van het universele krachtveld, de *conformer generator* van het Cambridge Crystallographic Data Center, en de *conformator*). Moloc vertoonde de hoogste graad van bemonsteringsefficiëntie en van volledigheid zonder duizenden conformaties of willekeurige ringsplitsing in twee halve lussen te produceren. Bovendien toonde Moloc het vermogen om zeer nauwkeurige bolvormige of platte conformaties te produceren op een manier die vergelijkbaar is met die van Prime, MacroModel en MD voor 75% van de bestudeerde dataset van macrocylische verbindingen. Deze bevindingen zullen verdere *in silico*-evaluaties van macrocylische verbindingen als potentiële bioactieve verbindingen mogelijk maken. Verder hebben we de kenmerken geïdentificeerd die verbeterd

moeten worden voor de ontwikkeling van nieuwe instrumenten voor de bemonstering en het ontwerp van macrocylische verbindingen, zoals die op basis van machinaal leren, om de conformaties van macrocylische verbindingen te voorspellen.

In Hoofdstuk 4 beschrijven we een gedetailleerd protocol voor achterwaartse koppeling ter identificatie van potentiële doelwitten voor 4-hydroxycoumarin (4-HC). Het moleculair koppelen is een nuttige en krachtige rekenmethode voor de identificatie van mogelijke interacties tussen kleine moleculen en farmacologische doelwitten. Het vermogen van een of enkele verbindingen om een grote dataset van eiwitten te binden, wordt *in silico* beoordeeld in achterwaartse koppeling. Deze strategie is nuttig voor het identificeren van moleculaire doelwitten van bioactieve 'weesverbindingen', het ontdekken van alternatieve medicijnindicaties (herbestemming) en het voorspellen van geneesmiddeltoxiciteit. Onze bevindingen laten zien dat RAC1 een doelwit is van 4-HC, wat helpt om enkele van de biologische effecten van 4-HC op kankercellen te verklaren. De strategie die in dit hoofdstuk wordt beschreven, kan gemakkelijk toegepast worden op andere verbindingen en datasets van eiwitten, om zo knelpunten in protocollen voor moleculaire koppelingen te overwinnen, met name in de aanpak via de achterwaartse koppeling.

Tot slot laat Hoofdstuk 5 zien hoe computationele methoden en experimentele resultaten gebruikt kunnen worden om een herbestemming te geven aan verbindingen die gebruikt zouden kunnen worden ter behandeling van COVID-19. De HCV NS3–4A proteaseremmers boceprevir en telaprevir werden in dit hoofdstuk ontdekt als 'hoofdprotease'(3CLpro)-remmers van SARS-CoV-2. We hebben ons gericht op de herbestemming van geneesmiddelen tegen SARS-CoV-2, omdat de COVID-19-gevallen snel aan het toenemen zijn, wat leidt tot een groeiend sterftecijfer en wat de wereldeconomie lamlegt. In tegenstelling tot een *de novo* geneesmiddelenontdekking, die er jaren over doet om van het concipiëren tot de preklinische fase en dan tot de commercialisering over te gaan, kan de herbestemming van medicijnen een kortetermijnoplossing vormen. Om die medicijnen te selecteren die hoogstwaarschijnlijk actief zijn tegen 3CLpro, hebben we een covalente koppelingsanalyse uitgevoerd. De geselecteerde geneesmiddelen werden beoordeeld in enzymremmingstests en co-gekristalliseerd, met de bedoeling om de hypothese te bevestigen dat α-ketoamide-geneesmiddelen covalent kunnen binden aan de cysteïne van de actieve site van de SARS-CoV-2 3CLpro. Tot slot hebben we vastgesteld dat boceprevir, en niet telaprevir, de replicatie in celkweek remt van SARS-CoV-2 en het muizenhepatitisvirus (MHV), een ander coronavirus. Volgens onze bevindingen zou het HCV-medicijn boceprevir klinisch getest moeten worden voor COVID-19 of andere coronavirusinfecties.

Kortom kunnen we zeggen dat deze hoofdstukken het belang, de toepassing en de beperkingen laten zien van computationele methoden in het state-of-the-artontwerpproces van geneesmiddelen.

Appendix

List of Publications

Acknowledgements

About the Author

List of Publications

First author (* indicates co-authorship)

1. **Angel J. Ruiz-Moreno**, Atilio Reyes Romero, Velasco-Velázquez, and Alexander Domling. *In silico design and selection of new tetrahydroisoquinoline-based CD44 antagonist candidates*. Molecules, 26(7), 1877, **2021**. DOI: 10.3390/molecules26071877.
2. Rick Oerlemans*, **Angel J. Ruiz-Moreno***, Yingying Cong, Nilima Dinesh Kumar, Marco A Velasco-Velazquez, Konstantinos Neochoritis, Jolanda Smith, Fulvio Reggiori, Matthew R Groves, Alexander Dömling. *Repurposing the HCV NS3-4A Protease Drug Boceprevir as COVID-19 Therapeutics*. RSC Med. Chem. **2021**. DOI: 10.1039/D0MD00367K.
3. Atilio Reyes Romero*, **Angel J. Ruiz-Moreno***, Matthew R. Groves, Marco Velasco-Velázquez, and Alexander Dömling. *Benchmark of Generic Shapes for Macrocycles*. J. Chem. Inf. Model. **2020**. DOI: 10.1021/acs.jcim.0c01038.
4. **Ruiz-Moreno A. J**., Dömling A., Velasco-Velázquez M. A. *Reverse Docking for the Identification of Molecular Targets of Anticancer Compounds*. In: Robles-Flores M. (eds) Cancer Cell Signaling. Methods in Molecular Biology, vol 2174. Humana, New York, NY. **2020**. DOI: 10.1007/978-1-0716-0759-6_4.
5. **Angel J. Ruiz-Moreno**, Atilio Reyes Romero, Constantinos Neochoritis, Marco Velasco-Velázquez, Matthew Groves, and Alexander Domling. *Gliptin Repurposing for COVID-19*. ChemRxiv, **2020**. DOI: 10.26434/chemrxiv.12110760.v1.
6. **Ruiz-Moreno AJ**, Torres-Barrera P, Velázquez-Paniagua M, Dömling Alexander, Velasco-Velazquez MA. *Guide for Selection of Relevant Cell Lines During the Evaluation of new Anti-Cancer Compounds*. Anti-Cancer Agents in Medicinal Chemistry. (8). p 1072 – 1081. **2018.** DOI: 10.2174/1871520618666180220120544.

Contributor

7. Afsaneh Sadremomtaz, Zayana M. AL-dahmani, **Angel J. Ruiz-Moreno,** Alessandra Monti, Chao Wang, Taha Azad, John Bell, Nunzianna Doti, Marco A. Velasco-Velázquez, Alexander Domling, Harry van Goor, Matthew R Groves. *Design and Biological Activities of Peptides that antagonize Angiotensin-Converting Enzyme-2 (ACE-2) interaction with the receptor binding spike protein of SARS-CoV-2*. ACS Journal of Medicinal Chemistry. July 30, 2010. DOI: 10.1021/acs.jmedchem.1c00477.
8. Jingyao Li, Vincenzo Di Lorenzo, Pravin Patil, **Angel J. Ruiz-Moreno**, Katarzyna Kurpiewska, Justyna Kalinowska-Tłuścik, Marco A. Velasco Velázquez, and Alexander Dömling. *Scaffolding-Induced Property Modulation of Chemical Space.* ACS Combinatorial Science, 2020. DOI: 10.1021/acscombsci.0c00072.
9. Inés Velázquez-Quesada, **Angel J. Ruiz-Moreno**, Diana Casique-Aguirre, Charmina Aguirre-Alvarado, Fabiola Cortés-Mendoza, Marisol de la Fuente-Granada, Carlos García-Pérez, Sonia M Pérez-Tapia, Aliesha González-Arenas, Aldo Segura-Cabrera, and Marco A Velasco-Velázquez. *Pranlukast Antagonizes CD49f and Reduces Stemness in Triple-Negative Breast Cancer Cells*. Drug Design, Development and Therapy. 14: 1799–1811. **2020.** DOI: 10.2147/DDDT.S247730.
10. Aguirre-Alvarado, C., Segura-Cabrera, A., Velazquez-Quesada, I., Hernandez-Esquivel, M. A., Garcia-Perez, C. A., Guerrero-Rodriguez, S. L., **Ruiz-Moreno A. J**., Rodriguez-Moreno, A., Perez-Tapia, S. M. and Velasco-Velazquez, M. A. *Virtual screening-driven*

repositioning of etoposide as CD44 antagonist in breast cancer cells. Oncotarget. 7 (17). p 23772-84. 2016. DOI: 10.18632/oncotarget.8180.

In preparation

11. Angel J. Ruiz-Moreno*, Mojgan Hadian*, Busra Ozdemir, Adhytia Mohan, Marco A. Velasco-Velazquez, and Alexander Dömling. *Structural review of IL-17A for the design of potential of novel inhibitors*. Manuscript in preparation.
12. André Boltjes*, Angel J. Ruiz-Moreno*, Markella Konstantinidou, Li Gao, John de Boer, Marco Velasco-Velazquez, and Alexander Dömling. *Diverse One-Pot Synthesis of Adenine Mimetics for Biomedical Applications.* Manuscript in preparation.
13. Angel J. Ruiz-Moreno*, Atilio Reyes Romero*, Katjee Knol, Sjoerd Wiarda, Marco A. Velasco-Velazquez, and Alexander Dömling. *A Protein Data Bank (PDB)based covalently targeted cysteine inhibitors database*. Manuscript in preparation.
14. Angel J. Ruiz-Moreno*, Atilio Reyes Romero*, Katjee Knol, Sjoerd Wiarda, Marco A. Velasco-Velazquez, and Alexander Dömling. *A Protein Data Bank (PDB)based covalently targeted histidine inhibitors database*. Manuscript in preparation.
15. Angel J. Ruiz-Moreno, Marco A. Velasco-Velazquez, and Alexander Dömling. *Implementation of computational methodologies for drug discovery.* Manuscript in preparation.

Acknowledgements

I'd like to thank the Universidad Nacional Autónoma de México (UNAM) and the University of Groningen (UG) for providing me with the incredible opportunity to pursue a double PhD program. Aside from all the people involved in the agreements, legal, and administrative matters required to make a collaboration between these two great universities a reality.

I would like to thank to the coordination of Programa de Doctorado en Ciencias Biomédicas, UNAM. Particularly to Dra. Aurea Orozco Rivas and Lic. Ivonne Torres Cortés, who patiently guided and supported me throughout the process of enrolling, developing, and graduating for the PhD at UNAM and UG in the context of my double degree agreement.

My heartfelt thanks go to my supervisors, Dr. Marco A. Velasco-Velázquez at UNAM and Prof. dr. Alexander Domling at UG.

Dear Marco, thank you for hosting me in your lab since my early days as a bachelor's student and for guiding me up to this point in my academic development. Since 2013, when we met for the first time, our interaction has been a path full of knowledge and grateful experiences. I remember one of your sentences to me in particular: "*At some point, students ended up performing in science as their academic parents.*" Hopefully, I was able to emulate your enthusiasm, wit, and ethics in your academic career. I am proud that the many lessons you taught me have already become a part of my scientific integrity and will continue to do so in the future. Experiment planning, presentation of results, writing reports and papers, but most importantly, the way I approach new projects and challenges with excitement are just a few but powerful examples. Thank you for your invaluable assistance, guidance, and encouragement throughout all stages of my professional and personal development.

Alex, I am grateful for the opportunity to have worked with you in the past years. I will be always grateful for your help and willingness to share your knowledge and ideas. Many of them led to incredible projects and collaborations. I'll never forget your constant encouragement to think outside the box, and to show me how to work independently. Which resulted in the development of some of my most valuable professional skills. You taught me how to study chemistry and biology rigorously while also having fun, being creative, and being smart.

Thanks to the members of my reading committees at RUG and UNAM who dedicate valuable time and expertise to comment on and review my thesis, which resulted in its final version: Prof. Carlos Camacho, Prof. Siewert-Jan Marrink, Prof. Gerrit J. Poelarends, Dra. Nuria Victoria Sánchez Puig, Dr. Martin González Andrade, Prof. Matthew Groves, and Dr. Enrique Ramon Ángeles Anguiano.

My gratitude goes out to everyone who took part in the collaborations and various projects. Thank you, Atilio, for your generosity and willingness to host me as a college during my first visit to Groningen. Your friendship resulted in several collaborations, all of which have pushed me ahead of my knowledge and skills while also contributing to my development as a scientist. I have fond memories of Christmas in Italy, as well as all the places we visited and enjoyed together. Best wishes and heartfelt consideration for all your current and future personal and professional endeavors. I wish you and to your family the biggest health and happiness.

Thank you, Rick, for your outstanding work in crystallization of SARS-CoV-2 3CLpro, not only in terms of quality but also in terms of speed. Our collaboration will serve as a reminder of the efforts and successes made during the corona pandemic.

As a computational scientist, I am aware that *"I am a hypothesis maker."* As a result, my deepest admiration and gratitude go to all the experts and students who contributed significantly to the experimental development required for the publication of the papers in this thesis. Yingying

Cong, Nilima Dinesh Kumar, Jolanda Smith, and Prof. Fulvio Reggiori, Patricia Torres-Barrera, Mireya Velázquez-Paniagua, Afsaneh Sadremomtaz, Zayana M. AL-dahmani, Alessandra Monti, Chao Wang, Taha Azad, John Bell, Nunzianna Doti, Prof. Harry van Goor, Vincenzo Di Lorenzo, Katarzyna Kurpiewska, and Justyna Kalinowska-Tłuści.

Special thanks to Prof. Matthew Groves and the UG Structural Biology group for inviting me to participate in several collaborations. Especially to Afsaneh, and Zayana. Aside from my supervisors, you placed your trust in my knowledge and abilities. I learned new techniques and challenged myself to broaden my collaborations as a result of it. Consequently, we were able to work on interesting projects together and produce very appealing results.

I'd like to express my heartfelt gratitude to all my friends/colleagues/mates who have spent time with me in the lab and whose interactions have made our work more pleasant, funny, and interesting.

First, I want to mention all the colleagues who became my friends at Molecular Pharmacology Lab at UNAM in Mexico, Charmina Aguirre, Sandra Guerrero, Fabiola Cortes, Abimael Mondragón, Luz Vázquez, Diana Pérez, Andrea Rodríguez, Luis Bahena, Ilselena Cortes y Alejandra Montes. One of my favorite memories is the lunch break, when we got together to talk about our projects, the latest news about TV shows, movies, and even philosophical topics. You all helped to make the lab feel like home with your friendship and affection. In a way, regardless of our personal or academic backgrounds, I believe we all have a deep love for Mexico and UNAM; for some of us, UNAM is and will always be our *alma mater*. Then, even though this thesis is written in English, raise your hand and shout the song of our illustrious university in my honor:

¡GOYA! ¡GOYA!
¡CACHUN, CACHUN, RA, RA!
¡CACHUN, CACHUN, RA, RA!
¡GOYA!
¡¡UNIVERSIDAD!!

Following that, I'd like to thank the wonderful and international staff of the Drug Design group. Patil, thank you for "adopting" me when I was completely lost in the synthesis lab and teaching me the joys of organic synthesis. I will always be amazed by your unique brilliance, talent, and passion. You take the nice chemistry and apply it in the kitchen to make the delectable masala. Thank you so much for all your help and kindness.

Angelina, Mojgan, Kumchok, Adithya, Shabnam It's been a pleasure to go out and talk with you. You have always been there to encourage me in difficult times, but also to celebrate with me and have fun. Your company has provided me with some of the most valuable experiences I've had in Groningen.

Hylke, Robin, André, Roberto Maryam, Marta, Markella, Fandi, Tryfonas, Dinos. It's been a pleasure to spend time with you in the lab, and I'm sorry that our interactions were so brief. Nonetheless, I am pleased that we worked together and had a good time. Thank you for your help and consideration. I wish you all a prosperous future.

Sara, Francesca, Federica, and Antonio are the Italian crew. Thank you for including me in your group and providing me with an entirely new perspective on Groningen and Europe. Going

out with you was an incredible and enjoyable experience. Your friendship inspired me to pursue my ambitions. I will be eternally grateful for your friendship.

The Chinese crew, Ruixue, Li, Zefeng, Biding, Qiang, Xin, Xiaofang, Qian, Jingyao, Qian. It's been a pleasure meeting you all. In my mind, I recall our time cycling together or eating hot pot at Zefeng's. I will remember you fondly and wish you all the best in your future endeavors. Zefeng, thank you for being such a great friend and office mate. When I was tired or frustrated, you always greeted me in the office or lab with a friendly smile or ready to crack a joke. The time spent at the office has never been dull. Thank you for your kindness and joy, Xin, my brother. You amaze me because, despite working so hard and for so long, you are always cheerful and upbeat. I admire you and hope you can continue to bring such joy to others.

Quiero agradecer a mis padres, quienes me han apoyado siempre en mi camino y decisiones a pesar de que estas nos han llevado a lugares diferentes. Muchas gracias por su amor, comprensión y constante consejo. Los llevo siempre conmigo y cada vez que las fuerzas se me agotan, recuerdo que cuento con ustedes y entonces tengo la voluntad para continuar de nuevo. Muchas gracias a mis hermanas por su cariño, solidaridad y amor. Me siento afortunado de haber crecido con ustedes y que a pesar de que cada uno ha elegido su camino, siempre podemos coincidir como hermanos para apoyarnos unos a otros. Soy afortunado de contar con ustedes. A mi hermano Cris. Estoy muy orgulloso de ti y verte persiguiendo tus sueños es siempre una motivación y recordatorio de que yo puedo hacer lo mismo. Me siento muy feliz por ti y espero que brilles en todo lo que hagas. Te admiro mucho por tu talento, pasión y sobre todo por tu sencillez y madurez. Los amo a todos.

Finalmente, quiero agradecer a todos aquellos que me han apoyado durante mi vida y que por alguna razón ya no están a mi lado. A pesar de que la vida nos allá separado, ya sea por muerte o por otra razón, no significa que haya olvidado su apoyo, amor, atención o cariño. Las invaluables experiencias y momentos los llevaré conmigo y por siempre serán parte de quien soy. Mi sincera gratitud.

Angel J. Ruiz-Moreno

Nov 2021

About the Author

Angel Jonathan Ruiz Moreno was born in Mexico City on April 1st, 1990. He received his Bachelor's degree in Pharmaceutical Chemistry from the Faculty of Chemistry at Universidad Nacional Autónoma de México (UNAM) in March 2015. He earned his Master's degree in Biochemistry from Instituto Politécnico Nacional (IPN) in Mexico on May 2015, under the supervision of Prof. Marco A. Velasco-Velázquez. Angel received training in immunopharmacology techniques for preclinical evaluation of anti-cancer compounds during his master's degree. In August 2017, he enrolled in a double PhD program at the UNAM and the University of Groningen (UG) in the Netherlands, motivated by his strong interest in drug design and computational biology under the guidance of Prof. Marco A. Velázquez of UNAM and Prof. Alexander Dömling of UG. The PhD research of Angel focused on computer-assisted design and synthesis of novel agents to combat diseases such as cancer and COVID-19, resulting in ten publications on high-impact journals, being first or co-first author in six of them. Angel has honed his skills in the application and use of computational methods in biology and chemistry. He also enjoys walking, reading, and writing in his spare time. Photography and interesting conversation are two of his favorite pastimes, and he is known for listening to music and signing while at home or at work.