



**Università  
degli Studi  
di Palermo**

AREA RICERCA E TRASFERIMENTO TECNOLOGICO  
SETTORE DOTTORATI E CONTRATTI PER LA RICERCA  
U. O. DOTTORATI DI RICERCA

Dottorato in Scienze Molecolari e Biomolecolari  
Dipartimento di Scienze e Tecnologie Biologiche, Chimiche e Farmaceutiche (STEBICEF)  
Settore Scientifico Disciplinare CHEM-07/A

Targeting the SARS-CoV-2 M<sup>PRO</sup> through covalent and non-covalent inhibition: molecular modeling and synthetic procedures for the development of new antiviral molecules

DOTTORE  
**ALESSIA BONO**

COORDINATORE  
**PROF.SSA GIOVANNA PITARRESI**

TUTOR  
**PROF. ANTONINO LAURIA**

CO TUTOR  
**PROF.SSA ANNAMARIA MARTORANA**

CICLO XXXVII  
ANNO CONSEGUIMENTO TITOLO 2024

# Table of Contents

<b>Abstract</b> .....	<b>4</b>
<b>1. Introduction</b> .....	<b>7</b>
1.1. SARS-CoV-2: Origins and Structure.....	8
1.2. Key targets for antiviral therapy against SARS-CoV-2 .....	9
1.3. SARS-CoV-2 M <sup>PRO</sup> : a druggable target.....	12
<b>2. Aim of the work</b> .....	<b>18</b>
<b>3. Results and Discussion</b> .....	<b>19</b>
<b>3.1. Targeting SARS-CoV-2 M<sup>PRO</sup> through Non-Covalent Inhibition</b> .....	<b>20</b>
3.1.1. Non-covalent dual binding inhibition strategy for SARS-CoV-2 M <sup>PRO</sup> .....	23
3.1.1.1. <i>In silico</i> ligand-based approach: DRUDIT.....	25
3.1.1.2. ADME properties.....	27
3.1.1.3. <i>In silico</i> structure-based studies: Molecular Docking at the catalytic site of SARS-CoV-2 M <sup>PRO</sup> .....	28
3.1.1.4. Statistical Analysis: Principal Component Analysis .....	33
3.1.1.5. <i>In silico</i> structure-based studies: Induced Fit Docking into the allosteric site of SARS-CoV-2 M <sup>PRO</sup> .....	36
3.1.1.6. Enzyme inhibition assay of SARS-CoV-2 M <sup>PRO</sup> .....	39
<b>3.2. Targeting SARS-CoV-2 M<sup>PRO</sup> through Covalent Inhibition</b> .....	<b>40</b>
3.2.1. Targeting Sars-Cov-2 Main Protease for the Treatment of COVID-19: Covalent inhibitors Structure-Activity Relationships insights and Evolution Perspectives .....	40
3.2.1.1. Carbonyl warhead .....	42
3.2.1.1.1. Aldehyde warhead.....	42
3.2.1.1.2. Ketone warhead .....	55
3.2.1.2. $\alpha$ -ketoamide warhead.....	58
3.2.1.3. Michael Acceptor group as warhead .....	63
3.2.1.4. Nitrile warhead .....	69
3.2.1.5. Ester warhead .....	72
3.2.1.6. Selenium/sulfur as electrophilic warhead: the case of <i>ebsele</i> n/ <i>eb</i> sulfur and analogues... ..	76
3.2.1.7. Electrophilic Warheads Reactivity.....	81
3.2.2. Covalent inhibition strategy through non-peptidomimetic inhibitors: Synthesis and <i>in vitro</i> exploration.....	83
3.2.2.1. Database design and synthesis .....	83
3.2.2.2. ADME properties.....	84
3.2.2.3. Enzyme inhibition assay of SARS-CoV-2 M <sup>PRO</sup> .....	86
3.2.2.4. Future Perspectives.....	87
3.2.3. Covalent inhibition strategy through peptidomimetic inhibitors: Computational Studies .....	88
3.2.3.1. <i>In silico</i> design of a new database of small molecules using CombiGlide® Schrödinger.....	89
3.2.3.2. Induced Fit Docking studies .....	94
3.2.3.3. DRUDIT and Principal Component Analysis .....	95
3.1.3.4. Covalent Docking Studies.....	98
<b>4. Conclusions and Future Perspectives: Computational Approaches in SARS-CoV-2 M<sup>PRO</sup> Inhibitors Discovery</b> .....	<b>104</b>
<b>5. Materials and Methods</b> .....	<b>107</b>
<b>5.1. <i>In silico</i> studies</b> .....	<b>107</b>
5.1.1. Ligand-based studies .....	107

5.1.1.1.	MOLDESTO: a new software for molecular descriptors calculation .....	107
5.1.1.2.	Biotarget Predictor Tool .....	107
5.1.2.	Combiglide: Design of a combinatorial library .....	108
5.1.2.1.	Reagent Preparation Panel .....	108
5.1.2.2.	Combinatorial Library Enumeration Panel .....	108
5.1.3.	Structure-based studies .....	108
5.1.3.1.	Ligand Preparation .....	108
5.1.3.2.	Protein Preparation .....	109
5.1.3.3.	Docking Validation .....	109
5.1.3.4.	Induced Fit Docking .....	110
5.1.3.5.	Covalent Docking .....	110
<b>5.2.</b>	<b>Multivariate Statistical Analysis .....</b>	<b>111</b>
<b>5.3.</b>	<b>Chemistry .....</b>	<b>112</b>
5.3.1.	General Information .....	112
5.3.2.	General procedure for the synthesis of 6-chloropyridin-2-yl benzoate (45a-m) .....	112
5.3.2.1.	Synthesis of 6-chloropyridin-2-yl 5-chloro-2-nitrobenzoate (45a) .....	113
5.3.2.2.	Synthesis of 6-chloropyridin-2-yl 2-amino-3,5-dichlorobenzoate (45b) .....	113
5.3.2.3.	Synthesis of 6-chloropyridin-2-yl 3,4,5-trimethoxybenzoate (45c) .....	114
5.3.2.4.	Synthesis of 6-chloropyridin-2-yl 4-methoxybenzoate (45d) .....	114
5.3.2.5.	Synthesis of 6-chloropyridin-2-yl 4-(trifluoromethyl)benzoate (45e) .....	114
5.3.2.6.	Synthesis of 6-chloropyridin-2-yl 4-methylbenzoate (45f) .....	114
5.3.2.7.	Synthesis of 6-chloropyridin-2-yl benzoate (45g) .....	115
5.3.2.8.	Synthesis of 6-chloropyridin-2-yl 4-(dimethylamino)benzoate (45h) .....	115
5.3.2.9.	Synthesis of 6-chloropyridin-2-yl 4-fluorobenzoate (45i) .....	115
5.3.2.10.	Synthesis of 6-chloropyridin-2-yl 4-bromobenzoate (45j) .....	115
5.3.2.11.	Synthesis of 6-chloropyridin-2-yl 2,3-dichlorobenzoate (45k) .....	116
5.3.2.12.	Synthesis of 6-chloropyridin-2-yl 4-chlorobenzoate (45l) .....	116
5.3.2.13.	Synthesis of 6-chloropyridin-2-yl 3-chloro-4-fluorobenzoate (45m) .....	116
<b>5.4.</b>	<b>Biological Procedures .....</b>	<b>116</b>
5.4.1.	Antiviral Assay .....	116
<b>6.</b>	<b><i>Appendix: Visiting research period at the University of Buckingham, Oxford Drug Design (October 2023 – May 2024), Supervisor: Prof. Paul W. Finn .....</i></b>	<b>118</b>
<b>6.1.</b>	<b>Machine learning model for predicting molecular activity using Molecular Descriptors and ElectroShape Descriptors .....</b>	<b>119</b>
<b>6.2.</b>	<b>Introduction .....</b>	<b>121</b>
6.2.1.	The role of Machine Learning and Artificial Intelligence in Drug Discovery .....	121
6.2.2.	Molecular Features: Molecular Descriptors and ElectroShape Descriptors .....	122
6.2.3.	Useful Benchmarking sets .....	123
6.2.3.1.	DUD and DUD-E datasets .....	124
6.2.3.2.	MUV dataset .....	125
6.2.3.3.	LIT-PCBA dataset .....	127
<b>6.3.</b>	<b>Aim of the project.....</b>	<b>128</b>
<b>6.4.</b>	<b>Results and Discussion.....</b>	<b>130</b>
6.4.1.	Molecular and ElectroShape Descriptors for MUV target sets.....	130
6.4.2.	Machine Learning Model Development .....	131
6.4.3.	Preliminary Results on MUV set – I ROUND .....	131
6.4.4.	Results on MUV set – II ROUND.....	136
6.4.5.	Top 25 Molecular Descriptors.....	142
6.4.6.	3-Fold Cross-Validation: Evaluating Model Performance with Top Descriptors.....	145
6.4.6.1.	Insights from 3-Fold Cross-Validation with Numeric Order Splitting .....	146
6.4.6.2.	Insights from 3-Fold Cross-Validation with Structural Identity Splitting.....	148
6.4.7.	Analysis of results across different Experimental scenarios.....	150

6.4.8.	Model Performance examination on LIT-PCBA dataset.....	152
<b>6.5.</b>	<b>Conclusions and Future Perspectives .....</b>	<b>157</b>
<b>6.6.</b>	<b>Materials and Methods .....</b>	<b>158</b>
6.6.1.	Benchmark Datasets.....	158
6.6.2.	Molecular Representation .....	158
6.6.3.	Features Selection .....	158
6.6.4.	Machine Learning Model Development .....	159
6.6.5.	Evaluation Metrics.....	159
6.6.6.	Structural Identity Splitting.....	159
6.6.6.1.	Step 1: Linkage Matrix Generation.....	159
6.6.6.2.	Step 2: Forming Flat Clusters .....	160
6.6.6.3.	Step 3: Limiting Cluster Formation.....	160
6.6.7.	Reciprocal Rank .....	160
6.6.8.	Validation on the LIT-PCBA Dataset.....	160
<b>7.</b>	<b>Bibliography.....</b>	<b>161</b>
	<b>List of Abbreviations .....</b>	<b>181</b>
	<b>Supplementary Material.....</b>	<b>184</b>
	<b>Supplementary Figures .....</b>	<b>184</b>
	<b>Acknowledgements .....</b>	<b>208</b>
	<b>Biography and CV of Alessia Bono .....</b>	<b>209</b>
	<b>List of peer reviewed publications.....</b>	<b>209</b>
	<b>List of conference proceedings.....</b>	<b>211</b>

## Abstract

**Background and rationale:** The COVID-19 pandemic has underscored the urgent need for specific pharmacological treatments beyond existing vaccines. One of the most attractive targets for antiviral therapies development is the SARS-CoV-2 Main Protease ( $M^{\text{PRO}}$ ), a key enzyme in viral life. The lack of  $M^{\text{PRO}}$  human homologs and its conservation rate among coronaviruses make this enzyme strategically important. Considering its mechanism of action,  $M^{\text{PRO}}$  inhibition could prevent the maturation of viral proteins and halt viral replication. X-ray crystallographic structures revealed that both catalytic and allosteric regions represent crucial sites for modulating its activity. The inhibition of the catalytic site of  $M^{\text{PRO}}$  is one of the most direct and effective strategies. Its high conservation across coronaviruses suggests that inhibitors could potentially offer broad-spectrum antiviral activity. On the other hand, allosteric inhibition presents a promising alternative strategy, targeting non-catalytic sites that are nevertheless crucial for enzyme function. Inhibition at these sites can occur through either covalent or non-covalent interactions.

*In silico* approaches are gaining increasing importance in drug development and clinical research. These techniques allow researchers to evaluate treatments qualitatively and quantitatively, leading to more practical and cost-effective experimentation. In the search for effective COVID-19 treatments, computational methods have become essential for discovering and developing SARS-CoV-2  $M^{\text{PRO}}$  inhibitors. Leveraging various computational methodologies accelerates the drug discovery process, reduces experimental costs, and enhances the precision of target identification and lead optimization.

**Results:** This PhD Thesis focuses on developing innovative *in silico* and synthetic protocols for identifying diverse anti-SARS-CoV-2 agents targeting  $M^{\text{PRO}}$ . To this purpose, three different inhibition strategies were explored: non-covalent inhibition, covalent inhibition through non-peptidomimetic inhibitors, and covalent inhibition through peptidomimetic inhibitors.

For non-covalent inhibitors, a series benzo[*b*]thiophene **2** and benzo[*b*]furan **3** compounds were identified using a hierarchical and hybrid virtual screening approach. We used the ligand-based Biotarget Predictor Tool (BPT), available in DRUDIT, to filter a large in-house structure database and identify small molecules with high affinity against the SARS-CoV-2  $M^{\text{PRO}}$  catalytic site. ADME properties were investigated through the SwissADME tool and docking studies confirmed DRUDIT prediction. Moreover, aiming at evaluating the possibility

of a dual binding mechanism of action, the identified hits were further investigated by means of statistical analysis and dockings studies into the dimerization site. Compounds **2i-l** also exhibited promising IC<sub>50</sub> when *in vitro* evaluated as inhibitors of the catalytic site, among them **2i** (IC<sub>50</sub> values of 70.4) and **2l** (IC<sub>50</sub> values of 45.9) are promising lead compounds for further development as antiviral agents with a dual binding activity.

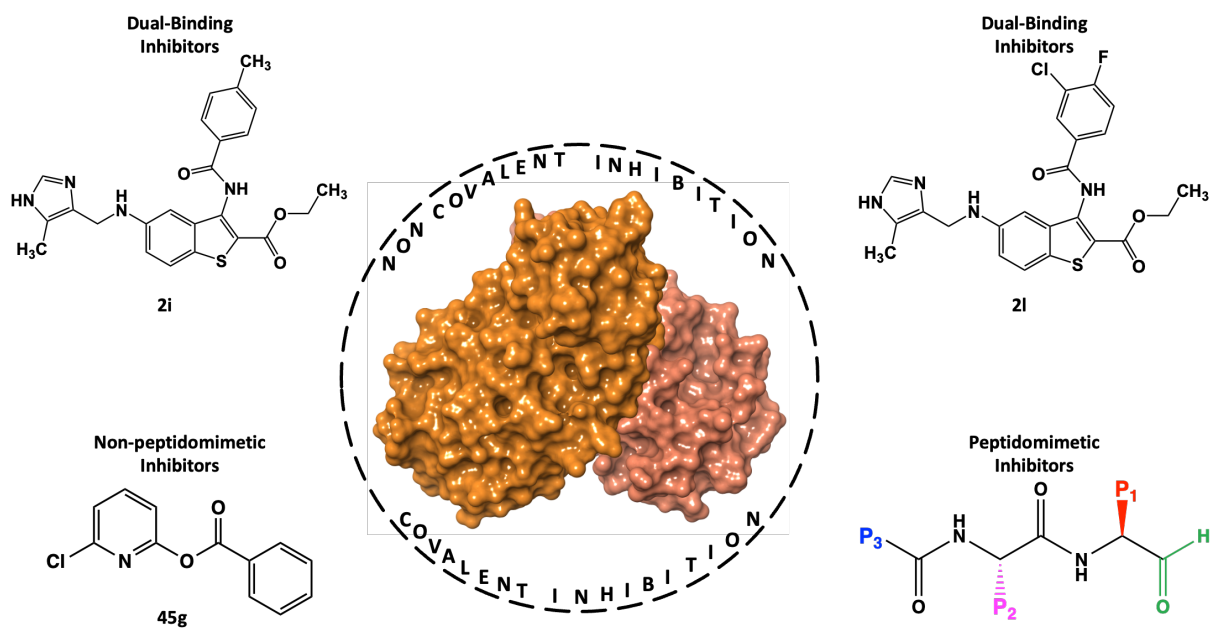
On the other hand, the covalent inhibition strategy has been extensively assessed through both peptidomimetic and non-peptidomimetic point of views.

In the case of non-peptidomimetic inhibitors, we focused on the rational design and synthesis of ester derivatives **45a-m**, with compound **45g** showing the most promise with an IC<sub>50</sub> value of 30 ± 6.6 μM. These esters offer the advantage of long-lasting inhibition, becoming a potent strategy for disrupting viral replication.

Additionally, a combinatorial library of 450 peptidomimetic compounds with aldehydic warheads was generated, refined to 388 compounds through docking studies, and further evaluated for covalent binding capabilities, revealing that compounds **57-62**, and **64** exhibited significantly higher affinities compared to known inhibitors, thus affirming the validity of the adopted design strategy.

**Conclusions:** Our findings demonstrate that integrating advanced computational tools offers a strategic and promising avenue for identifying new antiviral drugs. My thesis took the advantages of our in-house ligand-based tool, the BPT, available in DRUDIT, which allowed us to screen enormous ligands libraries. This tool integrated with both structure-based techniques and, interestingly, multivariate statistical analysis, has been applied to evaluate potential new SARS-CoV-2 M<sup>PRO</sup> inhibitors. This research project plays the groundwork for future research and the design of selective antiviral agents aimed at combating COVID-19, and supports ongoing efforts to combat SARS-CoV-2 and related coronaviruses.

## Graphical Abstract



## 1. Introduction

The Coronavirus Disease-19 (COVID-19) global pandemic has been etiologically caused by the highly infectious Severe Acute Respiratory Syndrome CoronaVirus 2 (SARS-CoV-2) [1,2]. SARS-CoV-2 is a  $\beta$ -coronavirus ( $\beta$ -CoV) that infects both animals and humans, causing respiratory diseases [3]. The virus infiltrates human cells via Spike (S) glycoproteins on its surface, which bind to the Receptor Binding Domain (RBD) of the Angiotensin-Converting Enzyme 2 (ACE2). ACE2 receptors are widely expressed in various pulmonary and extrapulmonary tissues, including the kidneys, heart, endothelium, and small intestine. This widespread distribution of ACE2 receptors suggests that, during COVID-19 infection, the virus can invade and damage not only the lungs but also other organs, leading to cardiovascular [4], thrombotic [5], pulmonary [6], and neurological complications [7].

The susceptibility to COVID-19 and its clinical manifestations are closely correlated with age [8]. In general, elderly patients (over 60 years) with comorbidities are more likely to develop severe respiratory illness that may require hospitalization or even result in death. A potential correlation between increased catabolism of the Nicotinamide Adenine Dinucleotide (NADH) cofactor and greater susceptibility to SARS-CoV-2 infection hypothesizes that higher NADH levels in younger populations might play a protective role against SARS-CoV-2 infection [9]. Indeed, most young people and children experience only mild illness (mild pneumonia) or are asymptomatic [10,11]. As of today, the World Health Organization (WHO) has reported over 775 million cases globally, with a mortality rate of approximately 1%.

In response to the outbreak, the scientific community has recommended stringent social distancing measures and the rapid development of vaccines to contain the spread of the virus. Over 20 vaccines have successfully reached clinical trials, demonstrating their efficacy in preventing severe COVID-19 [12]. However, considering the high mutation rate characteristic of this virus family, the long-term efficacy and safety of these vaccines remain under debate. This highlights the urgent need for more specific pharmacological treatments to complement vaccines in both the prophylaxis and treatment of COVID-19 [13]. Moreover, there is the need to develop effective therapeutic strategies against SARS-CoV-2. Over the past four years, recombinant proteins, monoclonal antibodies, immunological treatments, and drug repurposing have emerged as effective approaches based on comparative studies among coronaviruses [14-43].



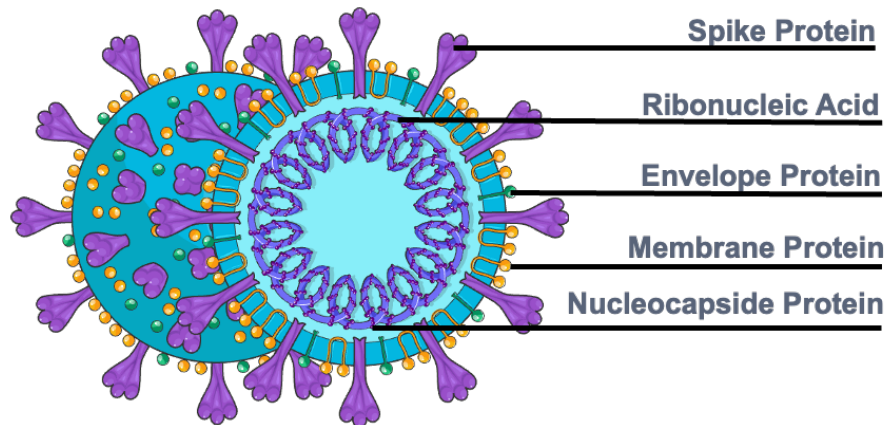
## 1.1. SARS-CoV-2: Origins and Structure

SARS-CoV-2 belongs to the family Orthocoronavirinae, a subfamily of viruses commonly known as coronaviruses, which are capable of causing a range of respiratory syndromes from mild to severe [3]. Among the viruses in the Coronavirus genus, those with the most significant impact on human health include SARS-CoV, Middle East Respiratory Syndrome CoV (MERS-CoV), and SARS-CoV-2. SARS-CoV was first reported in Asia in February 2003, though it is now known to have been present as early as 2002. The outbreak was contained by 2004. MERS-CoV emerged next, with its first identification in Saudi Arabia in 2012, followed by the emergence of SARS-CoV-2 in 2019 [44].

Phylogenetic, virological, epidemiological, and clinical analyses of the COVID-19 outbreaks in Wuhan, compared with those of SARS-2002 and MERS-2012, show that SARS-CoV-2 is phylogenetically similar to SARS-CoV, sharing approximately 79% of its genetic material. Additionally, SARS-CoV-2 exhibits about 50% genetic similarity with MERS-CoV and it is closely related to SARS-related coronaviruses identified in bats (SARS-related Bat CoVs). Consequently, these viruses share the same taxonomic classification, being part of the same order, family, subfamily, and genus [44,45].

Like all  $\beta$ -CoVs, SARS-CoV-2 has a genome composed of a single-stranded RNA of positive polarity, ranging from 27 to 30 kilobases (ssRNA+) [46]. This RNA is encapsulated within a helical capsid and is further surrounded by an envelope composed of a cellular-derived lipid bilayer. The genome of  $\beta$ -CoVs encodes four structural proteins: S, Envelope (E), Membrane (M), and Nucleocapside (N).

The virion, a spherical infectious viral particle (80-160 nm in diameter), features an outer phospholipid membrane (envelope) in which the S, M, and E proteins are embedded [47]. The genomic RNA is surrounded by phosphorylated N proteins within the lipid bilayer [48,49] (Figure 1). The SARS-CoV-2 genome comprises several components: a 5' Untranslated Region (5'UTR); an Open Reading Frame (ORF)1a/b encoding 16 non-structural proteins (NSPs); four genes coding for S, E, M, and N proteins; accessory genes coding for proteins ORF3a, ORF6, ORF7a, ORF7b, ORF8, and ORF10; a 3' Untranslated Region (3'UTR) [49].



**Figure 1.** SARS-CoV-2 structure with focus on its key structural proteins (S, M, E, and N).

## 1.2. Key targets for antiviral therapy against SARS-CoV-2

SARS-CoV-2 encodes several essential, structural, and non-structural, proteins that play vital roles in its replication and interaction with host cells. These viral proteins, together with some host proteins, represent potential targets for antiviral therapies due to their crucial functions in the virus life cycle. Below are the most important viral proteins that have been explored as therapeutic targets:

- **S proteins** are large glycoproteins that play a pivotal role in the virus ability to infect host cells. They are composed of two subunits, S1 and S2. The S1 subunit contains the RBD, which specifically binds to ACE2 [50]. The S2 subunit mediates membrane fusion between the viral and host cell membranes, allowing viral entry [51-53]. Due to their essential role in viral entry, S proteins are considering the primary target for neutralizing antibodies and are the basis for many COVID-19 vaccines, including mRNA and viral vector vaccines. For example, monoclonal antibodies like *casirivimab* and *imdevimab* neutralize the virus by binding to the RBD and blocking its interaction with ACE2 [54]. Furthermore, vaccines such as Pfizer-BioNTech and Moderna mRNA vaccines are designed to elicit an immune response specifically against the S protein. However, mutations in the S protein, especially in the RBD, can affect transmissibility and immune escape, making it a key focus of variant studies [49,55-57].
- **M proteins** are the most abundant structural protein in SARS-CoV-2 and are critical for shaping the viral envelope and facilitating the assembly of new virions. M proteins have a tripartite transmembrane domain that allows them to interact with other viral

proteins, including S and N proteins, during the viral assembly process. Although direct inhibitors of M proteins have not been developed yet, studies have shown that protease inhibitors, which block proteins involved in viral maturation, might indirectly influence M protein function [49,58].

- **N proteins** bind to the viral RNA genome, encapsulating it into a ribonucleoprotein complex, that is involved in several aspects of the viral replication cycle, including RNA transcription and replication. The N protein consists of two domains: an RNA-binding domain and a dimerization domain, which allow it to interact with viral RNA and other N proteins to efficiently package the genome. Beyond its structural role, the N protein is highly immunogenic, triggering strong immune responses in infected individuals. It is also implicated in modulating the host immune response, potentially aiding in viral immune evasion. Due to its conserved nature across coronaviruses, the N protein is often used as a target in diagnostic tests [49,59,60].
- **E proteins** are small but multifunctional proteins embedded in the viral envelope. They function as viroporins, forming ion channels that assist in the maturation and release of new viral particles. While no direct E protein inhibitors have been approved for clinical use, researchers have proposed that ion channel blockers, such as *amantadine*, might interfere with E protein function, though further studies are needed to confirm this approach [49,61].
- **TransMembrane PRotease, Serine 2 (TMPRSS2)** is a host protease that plays a vital role in facilitating SARS-CoV-2 entry into host cells. TMPRSS2 cleaves and activates the S protein on the surface of the virus, specifically at the S1/S2 cleavage site, enabling the fusion of the viral and host cell membranes. This proteolytic activation is crucial for the virus to gain entry into cells, as it allows the S protein to undergo conformational changes necessary for membrane fusion. TMPRSS2 is expressed in various tissues, including the respiratory tract, making it a key factor in the viral infection of lung epithelial cells. Inhibitors of TMPRSS2, such as *camostat mesylate* and *nafamostat*, have been investigated for their potential to block viral entry and reduce the severity of COVID-19 by preventing the priming of the S protein [50].
- **Papain-like protease (PL<sup>PRO</sup>)** is a NSP3 encoded by SARS-CoV-2, essential for both viral replication and immune evasion. Its primary function is to cleave specific sites in the

viral polyprotein, releasing other NSPs needed for the formation of the Replication-Transcription Complex (RTC), which is critical for viral RNA synthesis. In addition to its proteolytic role, PL<sup>PRO</sup> exhibits deubiquitinating and deISGylating activities. These functions allow the virus to counteract host immune responses by removing ubiquitin and Interferon-Stimulated Gene 15 (ISG15) from host proteins, thus impairing antiviral signaling pathways, including the type I interferon response. By modulating these pathways, PL<sup>PRO</sup> helps the virus evade immune detection. Given its dual role in both viral replication and immune modulation, PL<sup>PRO</sup> is considered a prime target for the development of antiviral drugs aimed at limiting viral spread and replication. Inhibitors like *GRL-0617* have shown promising in blocking both the proteolytic and immune-modulating activities of PL<sup>PRO</sup>, making them a valuable therapeutic strategy for limiting viral replication and immune evasion [62].

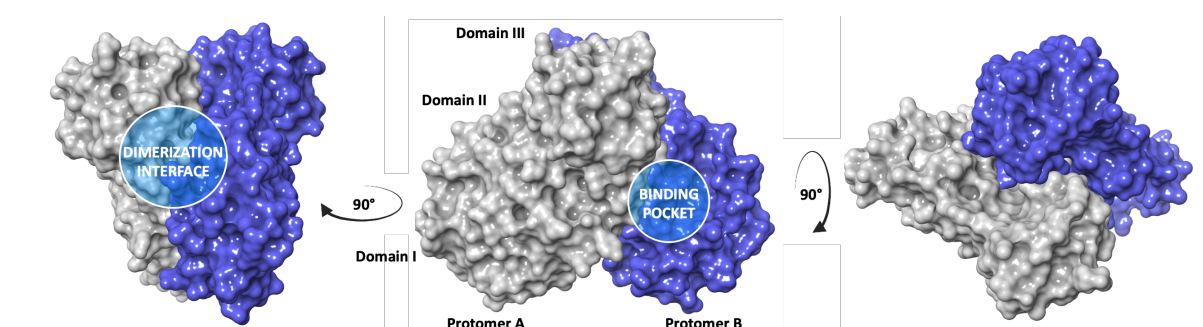
- **Main Protease (M<sup>PRO</sup>)**, also known as 3C-like protease (3CL<sup>PRO</sup>), is another crucial enzyme in the SARS-CoV-2 replication cycle, primarily involved in processing the viral polyproteins pp1a and pp1ab. By cleaving these polyproteins into functional NSPs, M<sup>PRO</sup> enables the assembly of the RTC, which is essential for viral replication. Unlike PL<sup>PRO</sup>, which also modulates host immune pathways, M<sup>PRO</sup> role is strictly proteolytic, focusing on viral protein maturation. M<sup>PRO</sup> is highly conserved among coronaviruses and functions as a homodimer, with its active site being a key target for antiviral drug development. The absence of similar proteases in human cells makes M<sup>PRO</sup> an ideal therapeutic target. Inhibition of M<sup>PRO</sup> has been shown to be a highly effective strategy in preventing the spread of SARS-CoV-2 and represents a cornerstone in the ongoing search for broad-spectrum antiviral therapies [63].

Among all the viral proteins targeted for antiviral therapy, M<sup>PRO</sup> stands out as the most critical for several reasons. First, its function is indispensable for the viral life cycle, as it processes the viral polyproteins pp1a and pp1ab into essential NSPs needed for replication and transcription. Unlike other proteases such as PL<sup>PRO</sup>, M<sup>PRO</sup> activity is strictly proteolytic, focused solely on viral protein maturation, making it highly specific. Its conservation across coronaviruses further enhances its value as a broad-spectrum target. Moreover, M<sup>PRO</sup> does not have human homologs, reducing the likelihood of OFF-target effects and making it an exceptionally safe target for drug development. Given these

factors, M<sup>PRO</sup> inhibitors have proven to be some of the most effective antivirals in clinical use against SARS-CoV-2. The blockade of M<sup>PRO</sup> halts viral replication without interfering with human proteins, ensuring both efficacy and safety in antiviral therapy. Thus, M<sup>PRO</sup> remains the most strategically important target in the development of therapies for not only SARS-CoV-2 but potentially other coronaviruses as well [63].

### 1.3. SARS-CoV-2 M<sup>PRO</sup>: a druggable target

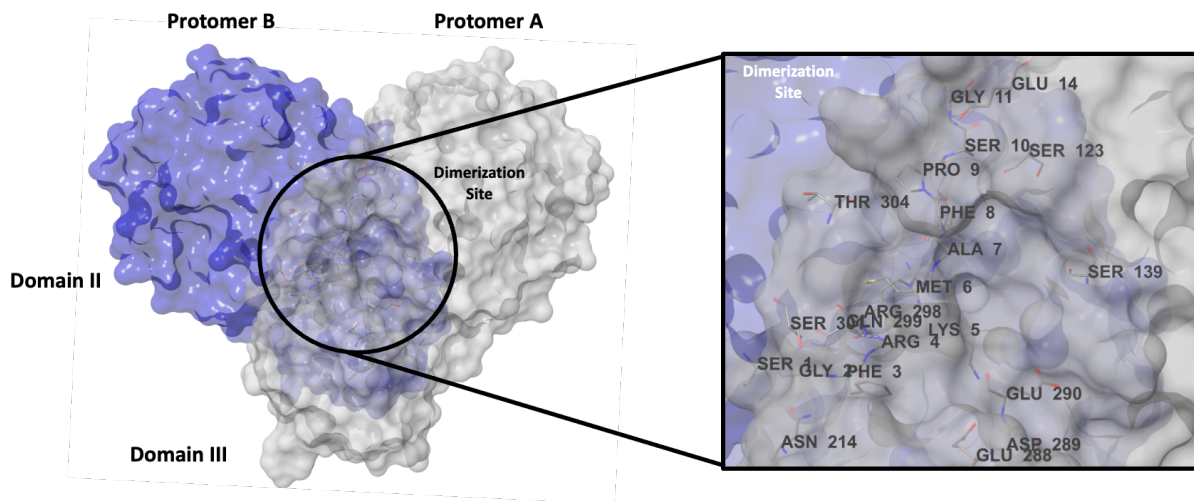
Structurally, M<sup>PRO</sup> is a homodimer and each protomer consists of 306 amino acids divided into 3 domains (I, II, and III): domain I (8-101) comprises 6  $\beta$ -strands and an  $\alpha$ -helix, while domain II (102-184) and domain III (201-303) include 6  $\beta$ -strands and 5  $\alpha$ -helices respectively. All domains are connected by long loop regions (Figure 2) [64].



**Figure 2.** X-Ray structure of dimeric SARS-CoV-2 M<sup>PRO</sup> [65].

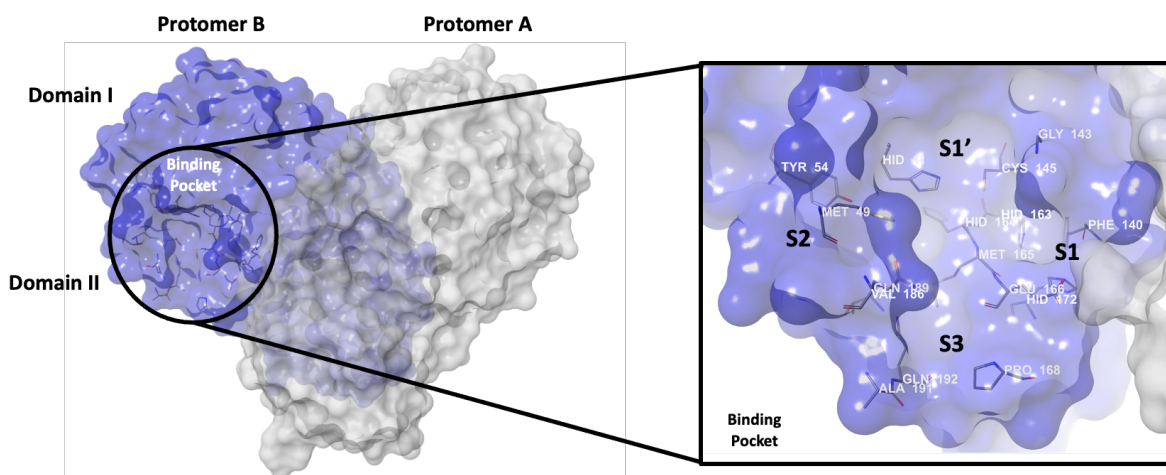
Dimerization of the protein is a necessary step for the catalytic activity, because of in the monomeric state the active site pocket collapses and is not available for the binding with the substrate. Detailed structural analysis of M<sup>PRO</sup> has revealed the presence of a potential allosteric site between domains II and III at the dimerization interface, which plays a crucial role in protease dimerization [66-68]. In particular, the *N*-finger residues Ser<sup>1</sup>-Gly<sup>11</sup>, residue Asn<sup>214</sup>, the region around residues Glu<sup>288</sup>, Asp<sup>289</sup>, Glu<sup>290</sup> in a tight contact with the *N*-finger, and the C-terminal last helix region around residues Arg<sup>298</sup>, Gln<sup>299</sup> are involved in the hydrogen bond interactions between the two SARS-CoV-2 M<sup>PRO</sup> protomers (Arg<sup>4</sup>/Glu<sup>290</sup>, Gly<sup>11</sup>/Glu<sup>14</sup>, Ser<sup>1</sup>/Glu<sup>166</sup>, Ser<sup>301</sup>/Ser<sup>139</sup>, Thr<sup>304</sup>/Glu<sup>166</sup>, Ser<sup>123</sup>/Arg<sup>298</sup>, Ser<sup>139</sup>/Gln<sup>299</sup>, Arg<sup>4</sup>/Gln<sup>299</sup>) [69], Figure 3.

Upon dimerization, a conformational change in M<sup>PRO</sup> activates the catalytic site. This structural reorganization stabilizes the active site orientation, thereby facilitating substrate binding and improving the enzyme catalytic efficiency. This newly stabilized orientation reveals the binding pocket, a crucial region where substrate molecules interact with the enzyme.



**Figure 3.** X-Ray structure of dimeric SARS-CoV-2 M<sup>PRO</sup> and focus on the dimerization site [65].

The binding pocket is organized into four subsites, S1', S1, S2, and S3/S4. In the region S1', between domains I and II, is located the catalytic site, characterized by the catalytic dyad (Cys<sup>145</sup> and His<sup>41</sup>). During the hydrolysis of the peptide bond, His<sup>41</sup> activates the nucleophilic -SH of Cys<sup>145</sup> by deprotonation, with subsequent stabilization of the adduct by the so-called "oxyanion hole", formed by the Gly<sup>143</sup> and Cys<sup>145</sup> backbones. The S1 region, characterized by the side chains of Phe<sup>140</sup>, His<sup>163</sup>, His<sup>164</sup>, Glu<sup>166</sup>, and His<sup>172</sup>, is highly specific for the glutamine residue [70]. The S2 consists of hydrophobic amino acids, such as Met<sup>49</sup>, Tyr<sup>54</sup>, Met<sup>165</sup>, Pro<sup>168</sup>, Val<sup>186</sup>, and finally S3/S4, which are particularly exposed to the solvent, involve Gln<sup>189</sup>, Ala<sup>191</sup>, Gln<sup>192</sup>, Gly<sup>251</sup> residues (Figure 4).



**Figure 4.** X-Ray structure of dimeric SARS-CoV-2 M<sup>PRO</sup> and focus on the catalytic site, with the four region S1, S1', S2 and S3/S4 highlighted [65].

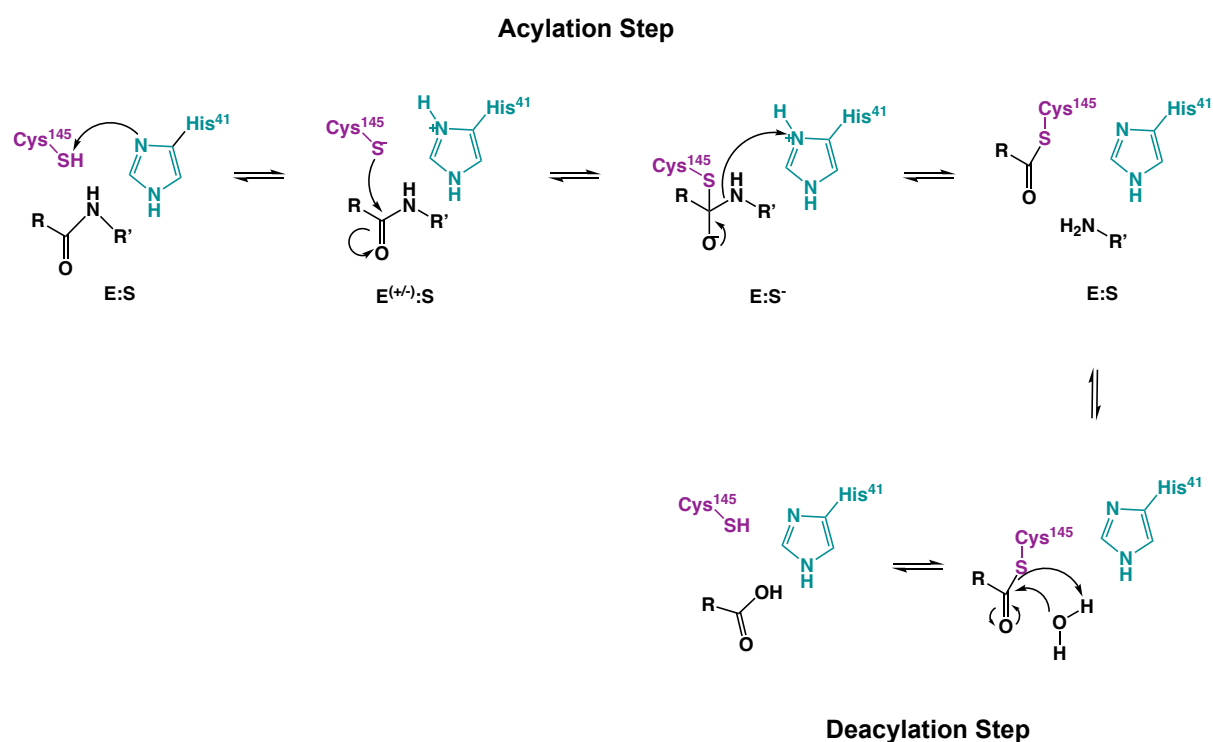
The M<sup>PRO</sup> mechanism of action involves particularly the S1' region. The catalytic dyad (His<sup>41</sup>/Cys<sup>145</sup>) is pivotal for the enzyme as it drives the process of peptide bond cleavage in the viral polyprotein. In this context, the enzyme-substrate initial complex corresponds to the neutral His<sup>41</sup>/Cys<sup>145</sup> dyad (**E:S** in Figure 5). From this stable state, the imidazole group of His<sup>41</sup> polarizes and acts as an activator by removing a proton from the thiol group (SH) of Cys<sup>145</sup> to form a highly nucleophilic CysS<sup>-</sup>/HisH<sup>+</sup> ion pair, that would react with the substrate (**E<sup>(+/-)</sup>:S** in Figure 5) [71]. The concomitant nucleophilic attack of CysS<sup>-</sup> on the carbonyl carbon atom of the peptide bond leads to a pseudo-stable intermediate (**E:S<sup>-</sup>** in Figure 5).

Then, the proton from the protonated HisH<sup>+</sup> is transferred to the N atom of the scissile peptide bond, forming an acyl-enzyme covalent intermediate (**E:S** in Figure 5).

After this acylation step, the recovery of the enzyme in the following deacylation stage is assisted by a water molecule activated by the His, which performs the hydrolysis of the thioester intermediate. This reaction results in the production of a smaller viral protein fragment with a new amino-terminal end and the regeneration of the enzyme for subsequent catalytic cycles.

The events that lead to the acylation step take place in a concerted but asynchronous way. The Transition State (TS) found for this mechanism is associated with the proton transfer from His<sup>41</sup> to the amide nitrogen atom of the peptide bond. The total free energy barrier associated with the acylation process, including the free energy cost of the ion pair formation, is 14.6 kcal·mol<sup>-1</sup>, while the free energy barrier associated with this first TS is 17.5 kcal·mol<sup>-1</sup>. Regarding the deacylation step, the free energy barrier is 15.6 kcal·mol<sup>-1</sup>. The second TS

observed during the deacylation corresponds to the separation of the Sulfur atom, and its free energy barrier is very close to the first one, 15.8 kcal·mol<sup>-1</sup>. Afterward, the leaving cysteine is stabilized by means of a proton transfer from the C-terminal group to the Sulfur atom, regenerating the enzyme in its more stable protonation state (a neutral catalytic dyad) and yielding the peptide fragment with a terminal unprotonated carboxylate [72,73].

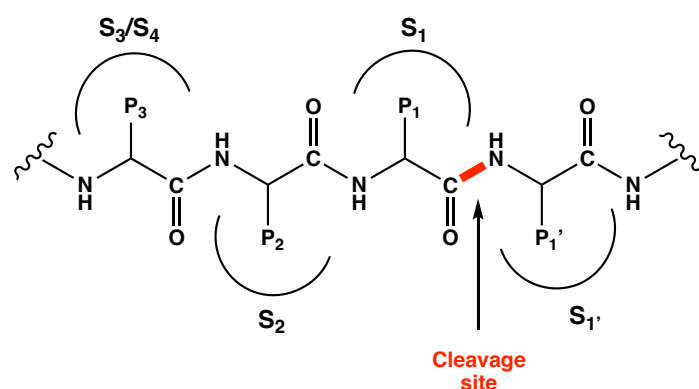


**Figure 5.** SARS-CoV-2 M<sup>PRO</sup> mechanism of action processed by the catalytic dyad (Cys<sup>145</sup> and His<sup>41</sup>, located in the S1' subregion of the binding pocket). Cys<sup>145</sup> is highlighted in purple, while His<sup>41</sup> in aqua green.

The strategic placement of His<sup>41</sup> and Cys<sup>145</sup> in the S1' region not only facilitates this reaction but also underscores the importance of this site in M<sup>PRO</sup> function. Disruptions in the S1' pocket could impede the enzyme ability to make timely and specific cuts in the polyprotein.

According to the Schechter and Berger nomenclature, the peptide substrate residues recognized by M<sup>PRO</sup> are referred to, from N to C terminus, P3-P2-P1-P1'-P2'-P3', with the hydrolysis taking place at the scissile bond between P1 and P1'. The multiple subsites S1', S1, S2, and S3 are labeled as they can recognize residues in P1', P1, P2, and P3, respectively. M<sup>PRO</sup> cleaves 11 of the 16 highly conserved recognition sites of the polyprotein pp1a/pp1ab, by specifically targeting the P1↓P1' positions (where ↓ denotes the peptide bond cleavage location, Figure 6).

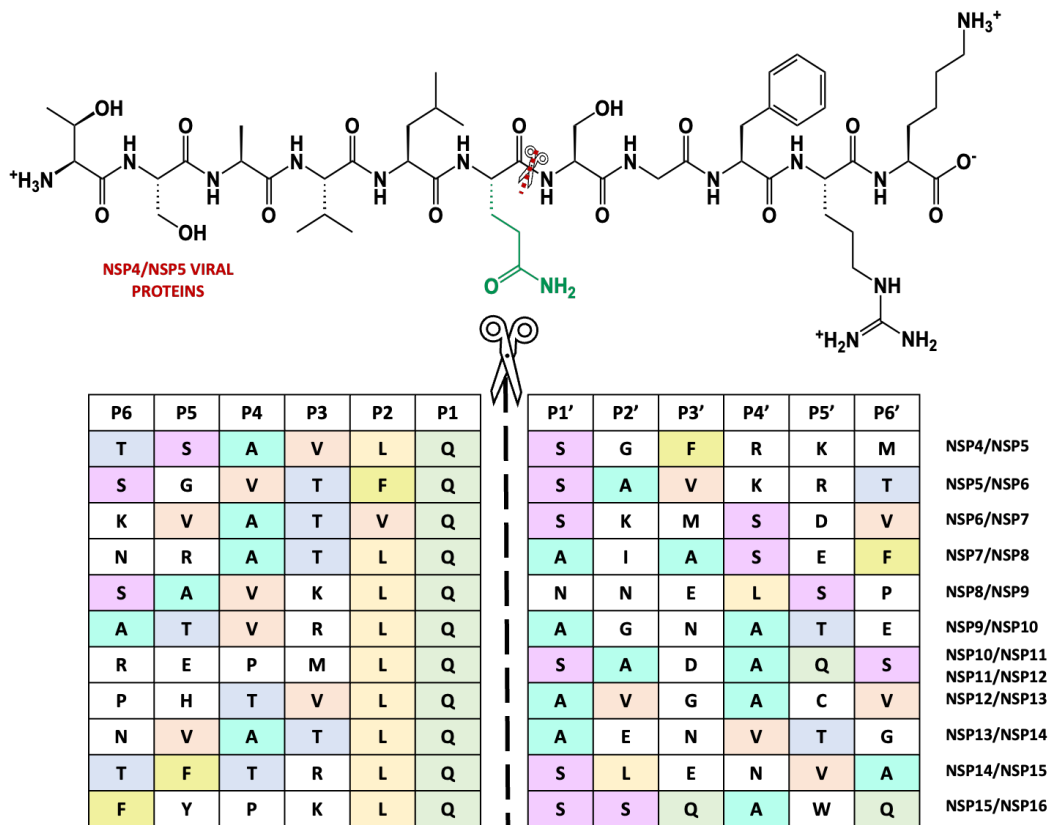




**Figure 6.** SARS-CoV-2 M<sup>PRO</sup> viral polyproteins. Sub-regions of the binding pocket (S<sub>1</sub>', S<sub>1</sub>, S<sub>2</sub>, and S<sub>3/S4</sub>) are labeled with S numbering complementary to the P fragments (P<sub>1</sub>', P<sub>1</sub>, P<sub>2</sub>, and P<sub>3</sub>) of the viral polyproteins.

Naturally, M<sup>PRO</sup> recognizes diverse cleavage sequences but with a stringent requirement of glutamine at P<sub>1</sub>, which gets recognized by the S<sub>1</sub> subsite in the active site of M<sup>PRO</sup>. S<sub>1</sub> subsite residues form an intricate network of interactions with the peptide substrate to ensure glutamine specificity at P<sub>1</sub> substrate position [74]. P<sub>2</sub> tolerates more hydrophobic amino acid with a clear preference for leucine, and P<sub>1</sub>' is a non-conserved prime recognition site (a small amino acid, such as serine, glycine, or alanine) [75,76]. Figure 7 shows the 2D structure of NSP4/NSP5 viral proteins and the P<sub>6</sub>-P<sub>5</sub>-P<sub>4</sub>-P<sub>3</sub>-P<sub>2</sub>-P<sub>1</sub>-P<sub>1</sub>'-P<sub>2</sub>'-P<sub>3</sub>'-P<sub>4</sub>'-P<sub>5</sub>'-P<sub>6</sub>' sequences for NSP4-NSP16.

Since P<sub>4</sub>, P<sub>3</sub>, P<sub>2</sub>, P<sub>1</sub>, and P<sub>1</sub>' residues interact with the substrate recognition sites (S<sub>3/S4</sub>-S<sub>1</sub>'), a recurrence in amino acids is appreciable in contrast with a higher variation rate for P<sub>2</sub>', P<sub>3</sub>', and P<sub>4</sub>'. In details, the analysis of amino acids in sequences highlights that residues as Ser and Ala are commonly found in P<sub>1</sub>', confirming that small aliphatic residues are favored at this position. The recurrent glutamine residue in P<sub>1</sub> strongly suggests that Gln is critical for substrate binding and cleavage; a mutation of Gln would abolish the substrate activity completely. P<sub>2</sub> position requires a large hydrophobic residue (mostly Leu, but also Phe and Val), indeed, hydrophobic interactions between the P<sub>2</sub> residue and the S<sub>2</sub> pocket of the enzyme are crucial for substrate recognition. As well as P<sub>2</sub>, P<sub>3</sub> residues must be characterized by substituent groups capable of favorable hydrophobic interactions with the amino acids of S<sub>3/S4</sub>. The preferred amino acids in P<sub>3</sub> are Val, Thr, Met, Lys, and Arg.



**Figure 7.** 2D structures of the viral NSP4 and NSP5 proteins are depicted alongside the P6-P5-P4-P3-P2-P1-P1'-P2'-P3'-P4'-P5'-P6' cleavage sequences for NSP4 to NSP16. The image highlights the recurrent glutamine residues at the P1 position, a critical feature across these sequences.

Since no human host-cell proteases are known with this substrate specificity, M<sup>PRO</sup> inhibition could drive to blockage of viral replication [49], making itself the most attractive target for COVID-19 antiviral treatment.

## 2. Aim of the work

In response to the unprecedented global health crisis posed by the COVID-19 pandemic, researchers worldwide focused on the need of effective antiviral treatments, with the SARS-CoV-2 M<sup>PRO</sup> emerging as a critical target for therapeutic intervention due to its essential role in the viral replication cycle. In this context, the aim of this PhD thesis is to develop and propose innovative *in silico* and synthetic protocols for the identification and optimization of diverse anti-SARS-CoV-2 agents specifically targeting M<sup>PRO</sup>. Specifically, the research focuses on three key approaches:

1. Non-Covalent Inhibition: This approach focuses on the design of small molecules that can reversibly bind to the catalytic and allosteric sites of M<sup>PRO</sup>. Using hierarchical and hybrid virtual screening techniques, we aim to discover potent inhibitors that can effectively disrupt enzyme activity with a dual binding mechanism of action.
2. Covalent Inhibition through Non-Peptidomimetic Inhibitors: In this strategy, we aim to rationally design non-peptidomimetic compounds, with an ester moiety capable of forming covalent adducts with the catalytic Cys<sup>145</sup> of M<sup>PRO</sup>, thereby achieving long-lasting inhibition. The development and optimization of these non-peptidomimetic inhibitors aim to enhance their potency and efficacy against the virus.
3. Covalent Inhibition through Peptidomimetic Inhibitors: This strategy emphasizes the identification of peptidomimetic inhibitors, with an aldehydic electrophilic warhead, that mimic the natural substrates of M<sup>PRO</sup>, ensuring specific interactions that facilitate effective inhibition.

Through *in silico* methodologies, synthetic strategies, and *in vitro* evaluations, this work aspires to contribute to the discovery and development of effective antiviral agents against COVID-19 and related coronaviruses.

### 3. Results and Discussion

Given the essential role of M<sup>PRO</sup> in viral replication and the absence of a human homolog, this protease has emerged as a prime target for antiviral drug development. Inhibiting the catalytic site of M<sup>PRO</sup> is one of the most direct and effective strategies for halting viral replication. Inhibitors that bind to this active site can effectively block the enzyme catalytic activity, thereby preventing the maturation of viral proteins and halting viral replication. The high conservation of the catalytic site across different coronaviruses suggests that targeting it not only provides a means to inhibit SARS-CoV-2 but also opens the possibility for broad-spectrum antiviral activity against other coronaviruses. However, potential challenges such as the development of resistance and the specificity of inhibitors highlight the need for ongoing efforts in optimizing these catalytic site inhibitors. Continued research into the structure and dynamics of the catalytic site is crucial for refining these inhibitors and ensuring their long-term efficacy [13,49,63,70,76].

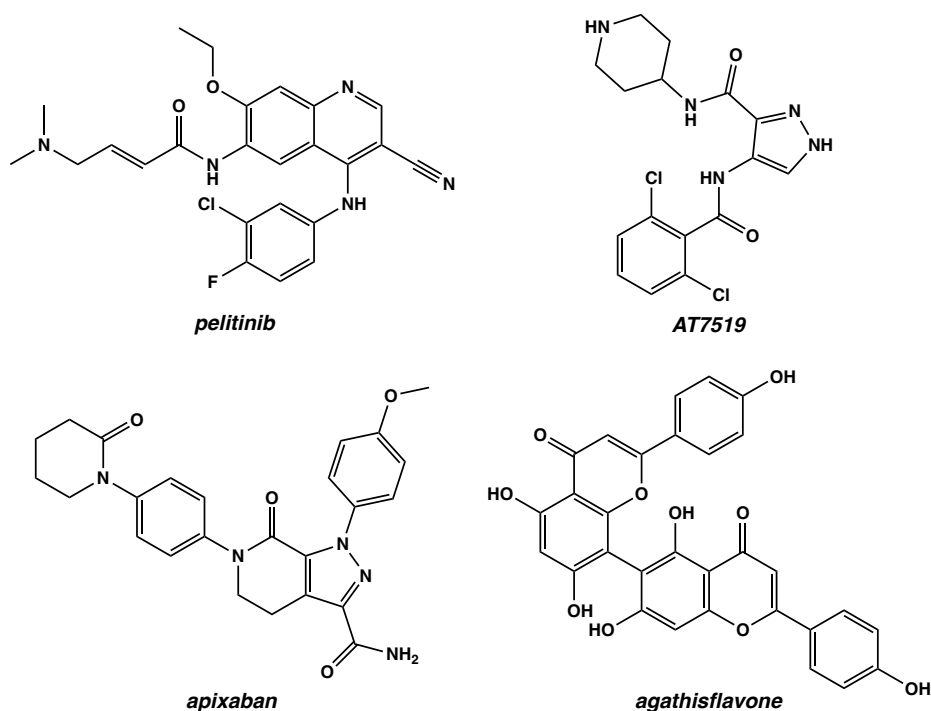
In addition to catalytic site inhibition, allosteric inhibition presents a promising alternative strategy in antiviral drug development. This approach targets non-catalytic sites that are nevertheless crucial for enzyme function. In the case of SARS-CoV-2 M<sup>PRO</sup>, several allosteric pockets on its surface have been identified as essential for its catalytic activity. These sites are in areas distal to the main catalytic pocket, including at the dimerization interface [66,77,78]. Some of the most notable allosteric sites are the distal site and the dimerization site, located at the interface between domains II and III. The distal site is characterized by key contacts such as Arg<sup>131</sup>-Thr<sup>199</sup>, Arg<sup>131</sup>-Asp<sup>289</sup>, Pro<sup>132</sup>-Thr<sup>196</sup>, Asp<sup>197</sup>, Thr<sup>198</sup>-Asn<sup>238</sup>, and Tyr<sup>239</sup>-Leu<sup>287</sup> [79], and is surrounded by residues forming  $\beta$ -sheets and  $\alpha$ -helices; while the dimerization site is involved in the interactions between the two SARS-CoV-2 M<sup>PRO</sup> protomers, making it an attractive target for allosteric inhibitors [68]. Computational studies predict that targeting these allosteric sites could result in modulation of M<sup>PRO</sup> activity by inducing conformational changes or altering thermal fluctuations around a fixed, mean conformation [80]. Given the challenges associated with targeting the active site directly, focusing on distal and dimer allosteric sites could broaden our understanding of these regions as molecular targets, ultimately contributing to the development of novel broad-spectrum inhibitors of SARS-CoV-2 M<sup>PRO</sup> [81].

Inhibition at these sites can occur through either covalent or non-covalent interactions. Covalent inhibitors, which form long-lasting time bonds, can achieve sustained inhibition but may pose a risk of OFF-target effects and toxicity due to their permanent binding. On the other hand, non-covalent inhibitors bind reversibly, allowing for more controlled modulation of enzyme activity. This reversible nature can reduce the likelihood of adverse effects and offer greater flexibility in dosage. The strategic balance between efficacy and safety makes the choice between covalent and non-covalent inhibitors a critical consideration in the design of antiviral therapies [82].

### 3.1. Targeting SARS-CoV-2 M<sup>PRO</sup> through Non-Covalent Inhibition

Non-covalent inhibition of SARS-CoV-2 M<sup>PRO</sup> represents a strategic approach in antiviral drug design, leveraging reversible interactions to modulate the protease activity. Examples of non-covalent inhibitors targeting SARS-CoV-2 M<sup>PRO</sup> include those that bind either to the dimerization site or the catalytic site.

*Pelitinib* (Figure 8), an anticancer drug developed as an Epidermal Growth Factor Receptor (EGFR) inhibitor, has been shown to form a complex with the SARS-CoV-2 M<sup>PRO</sup> dimerization site, with an EC<sub>50</sub> = 1.25 μM, indicating its efficacy in disrupting the protease dimerization [83]. *AT7519* (Figure 8), an inhibitor of Cyclin-Dependent Kinases (CDKs), binds at the M<sup>PRO</sup> dimer interface, specifically involving domains I and II [83,84]. Currently under investigation as an anticancer drug, *AT7519* effectively interferes with protein dimerization, thereby inhibiting the catalytic function of M<sup>PRO</sup>. Additionally, *apixaban* (Figure 8), a Food and Drug Administration (FDA)-approved oral anticoagulant for treating thromboembolic disease, inhibits M<sup>PRO</sup> by binding to the allosteric site via a non-competitive inhibition mechanism [85]. Furthermore, *agathisflavone* (Figure 8), a bioflavonoid, has emerged as a potential non-competitive M<sup>PRO</sup> inhibitor with an EC<sub>50</sub> of 4.32 μM, highlighting its promising antiviral activity [86]. These findings underscore the potential of allosteric inhibition as a complementary strategy to traditional catalytic site targeting, broadening the scope for antiviral drug development against SARS-CoV-2.



**Figure 8.** Chemical structures of SARS-CoV-2 M<sup>PRO</sup> allosteric inhibitors.

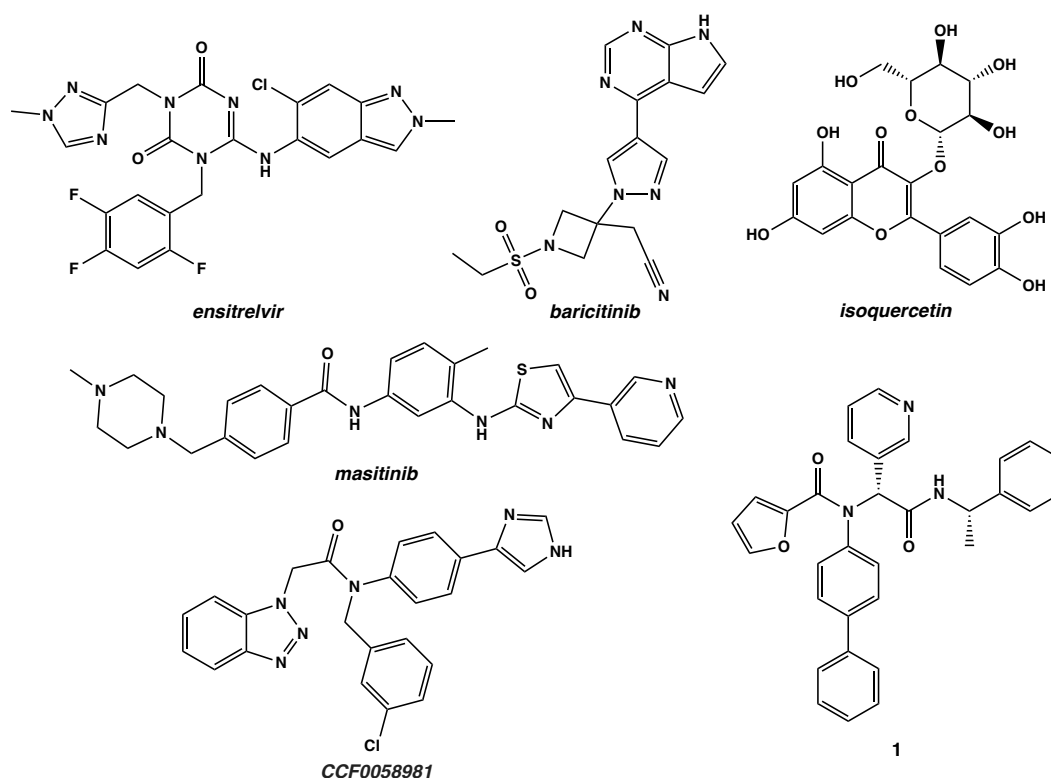
Figure 9 illustrates the 2D structures of several promising non-covalent M<sup>PRO</sup> inhibitors of the catalytic site. *Ensitrelvir* (*Xocova*, Figure 9), the first clinically approved non-peptide competitive inhibitor, targets the M<sup>PRO</sup> active site through non-covalent interactions [87,88]. With an IC<sub>50</sub> value of 13 nM and an EC<sub>50</sub> value of 0.37 μM in SARS-CoV-2 infected Vero E6 cells, *ensitrelvir* received the Emergency Use Authorization (EUA) in the US and Japan following its success in phase 2/3 clinical trials (JPRN-jRCT2031210350) [89].

Similarly, *baricitinib* (Figure 9), an FDA-approved oral tyrosine kinase inhibitor marketed as *olumiant*, acts as a competitive small molecule inhibitor of SARS-CoV-2 3CL<sup>PRO</sup> with an IC<sub>50</sub> in the micromolar range. It has also been granted EUA by the FDA for hospitalized COVID-19 patients [90,91].

*Masitinib* (Figure 9), another tyrosine kinase inhibitor, demonstrated an IC<sub>50</sub> of 2.5 μM against SARS-CoV-2 M<sup>PRO</sup> and an EC<sub>50</sub> of 3.2 μM in SARS-CoV-2 infected A549/ACE2 cells, confirming non-covalent binding to the catalytic pocket [92]. Currently in phase 3 clinical trials (ClinicalTrials.gov Identifier: NCT05441488) and a candidate for treating progressive forms of multiple sclerosis [93], *masitinib*, in combination with *isoquercetin* (Figure 9), has shown effectiveness for the early treatment of COVID-19 (ClinicalTrials.gov Identifier: NCT04622865).

Compound **1** (Figure 9), synthesized through the one-pot Ugi-4CR method, exhibited an  $IC_{50}$  value of  $0.20 \mu\text{M}$  against SARS-CoV-2  $M^{\text{PRO}}$  and an  $EC_{50}$  value of  $1.27 \mu\text{M}$  against SARS-CoV-2 in Vero E6 cells.

Additionally, *CCF0058981* (Figure 9), derived from a SARS-CoV-1 inhibitor, has been identified as a non-covalent competitive inhibitor of SARS-CoV-2  $3\text{CL}^{\text{PRO}}$ , with an  $IC_{50}$  value of  $68 \text{ nM}$  and potent antiviral activity with  $EC_{50}$  values of  $0.497 \mu\text{M}$  in SARS-CoV-2 infected Vero E6 cells [94,95].



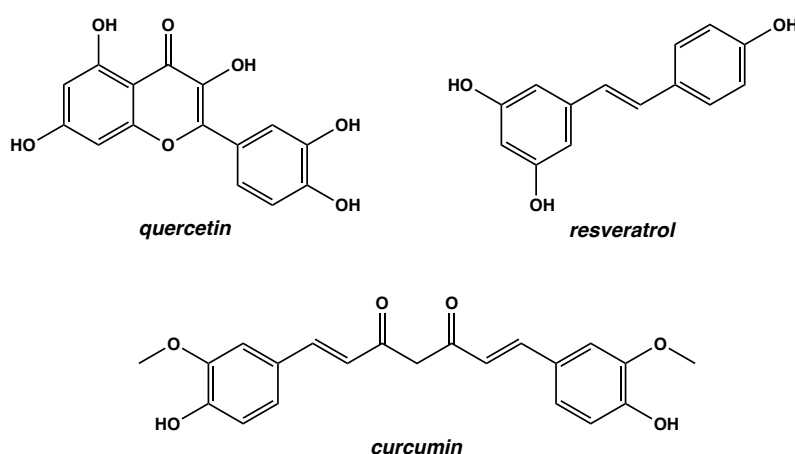
**Figure 9.** Chemical structures of SARS-CoV-2  $M^{\text{PRO}}$  non-covalent inhibitors of the catalytic site.

In addition to synthetic inhibitors, natural compounds represent another promising class of non-covalent competitive inhibitors of the  $3\text{CL}^{\text{PRO}}$  enzyme, with potential applications in treating viral infections and enhancing the host immune response [96-100].

Natural products like *quercetin*, *resveratrol*, and *curcumin* (Figure 10) have demonstrated notable therapeutic benefits in previous coronavirus outbreaks such as SARS-CoV and MERS-CoV. *Quercetin*, one of the most potent competitive inhibitors of  $3\text{CL}^{\text{PRO}}$ , has an  $IC_{50}$  value of  $7.40 \mu\text{M}$  and is currently undergoing clinical trials (ClinicalTrials.gov Identifier: NCT04861298) for early-stage and mild-to-moderate symptomatic COVID-19 outpatients [101-103].

Moreover, *quercetin* has received FDA Generally Recognized As Safe (GRAS) status for use as a dietary supplement [102,103].

*Resveratrol*, a polyphenol known for its 3CL<sup>PRO</sup> inhibitory activity *in vitro*, has also shown effectiveness against SARS-CoV-2 in cell culture assays [99,104,105]. In a clinical trial (ClinicalTrials.gov Identifier: NCT04400890), *resveratrol* demonstrated a reduction in the incidence of COVID-19-related hospitalizations, emergency room visits, and pneumonia [106]. Recent dose-response studies have revealed that *curcumin* binds to the 3CL<sup>PRO</sup> active site with low micromolar IC<sub>50</sub> values, highlighting its potential as a therapeutic agent against SARS-CoV-2 [99,107]. The activity of these natural compounds underscores their potential role in both the prevention and treatment of COVID-19, complementing conventional antiviral therapies.



**Figure 10.** Chemical structures of natural compounds.

In conclusion, the non-covalent inhibition of SARS-CoV-2 M<sup>PRO</sup>, whether targeting the catalytic or allosteric sites, has proven to be a versatile approach, with several molecules effectively modulating enzymatic activity and disrupting the viral life cycle. In light of these findings, it would be particularly interesting to explore a dual-binding inhibition, aiming to discover molecules capable of inhibiting the protease both competitively at the active site and allosterically, offering a potentially more robust strategy for antiviral drug development.

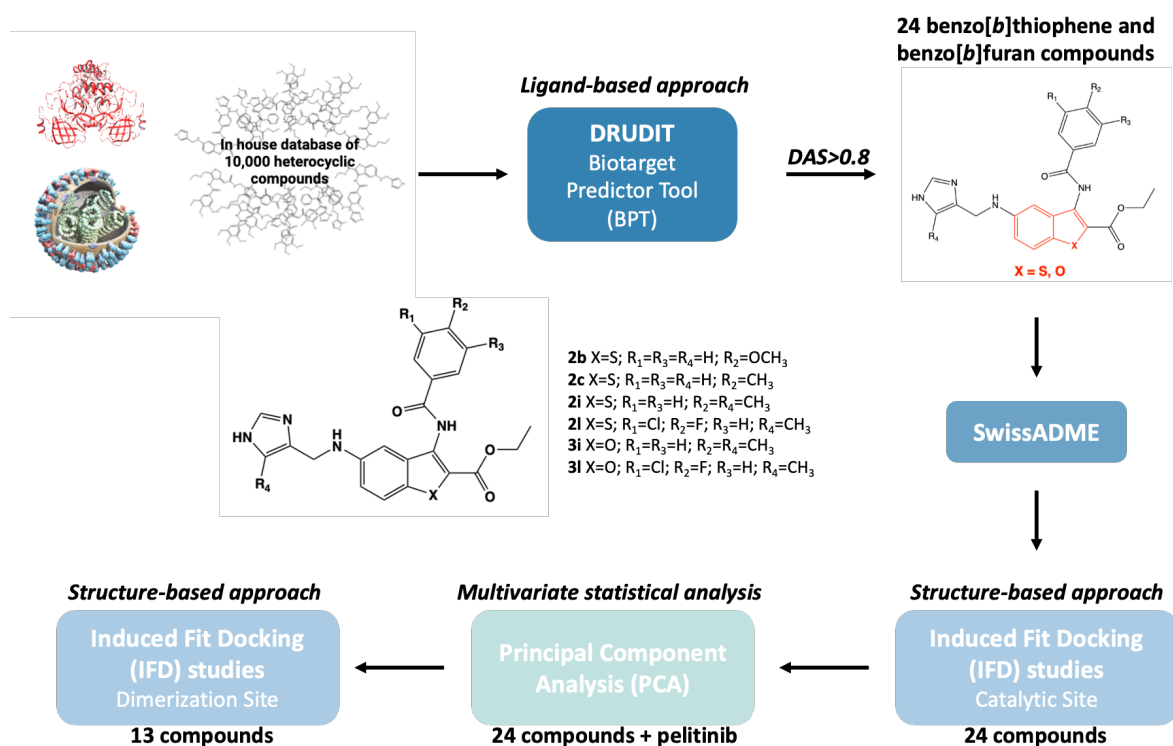
### 3.1.1. Non-covalent dual binding inhibition strategy for SARS-CoV-2 M<sup>PRO</sup>

The search for effective antiviral candidates against SARS-CoV-2 has become a critical area of research, with virtual screening emerging as a reliable approach. Inhibition of either the catalytic site or the dimerization interface of M<sup>PRO</sup>, which has proven to be one of the most



attractive viral targets, represents a promising approach to disrupting the viral life cycle. Targeting the dimerization of M<sup>PRO</sup> offers an alternative strategy by preventing the formation of the functional dimer, thereby impeding the protease ability to process viral polyproteins. Disrupting dimerization not only affects the enzyme activity but also offers a complementary approach to catalytic inhibition, potentially enhancing the overall antiviral efficacy. Together, these dual strategies—targeting both the catalytic site and the dimerization interface—provide a multifaceted approach to inhibiting M<sup>PRO</sup>, with the potential to yield robust antiviral agents against SARS-CoV-2.

We conducted an *in silico* analyses to investigate the inhibitory activity of small molecules against SARS-CoV-2 M<sup>PRO</sup>, aiming to identify new dual binding site modulators that could serve as promising non-covalent antiviral agents. The methodology employed is illustrated in Figure 11, which outlines the mixed ligand-structure virtual screening process used for the identification of potential SARS-CoV-2 M<sup>PRO</sup> inhibitors.



**Figure 11.** *In silico* protocol workflow for the identification of new SARS-CoV-2 M<sup>PRO</sup> inhibitors as effective antiviral molecules in COVID-19 treatment.

Our computational protocol was structured in a hierarchical manner, beginning with a ligand-based filtering phase that utilized DRUDIT (DRUG Discovery Tools, an open access web-service,

accessible at [www.drudit.com](http://www.drudit.com)) [108]. This initial step allowed us to filter small molecules based, leading to the selection of two clusters of compounds with known antitumor activity, characterized by benzothiophene or benzofuran scaffolds [108-110].

Further refinement was achieved by assessing the Absorption, Distribution, Metabolism, and Excretion (ADME) properties using the SwissADME web tool, which supports the discovery of biocompatible drug candidates [111]. Subsequently, the *in silico* affinity of 24 selected compounds was evaluated through structure-based molecular docking studies at the catalytic binding site of SARS-CoV-2 M<sup>PRO</sup>.

To enhance the robustness of the screening, we integrated the molecular descriptors of these 24 compounds with those of the non-covalent allosteric inhibitor *pelitinib*, allowing for a comprehensive multivariate analysis. Ultimately, 13 benzothiophene and benzofuran derivatives were identified as potential inhibitors and further analyzed using structure-based docking protocols targeting the dimerization site of SARS-CoV-2 M<sup>PRO</sup>.

#### **3.1.1.1. *In silico* ligand-based approach: DRUDIT**

To perform the *in silico* ligand-based approach, we utilized the DRUDIT, a resource created by the research group to support drug discovery, which includes the developed Biotarget Predictor (BPT) and Antiproliferative Activity Predictor (APP) tools. The main features of DRUDIT are its ease of use and accessibility to researchers worldwide.

The proposed tools are based on molecular descriptors, calculated through the MOLEcular DEscriptor TOol (MOLDESTO), our proprietary software, capable of generating over 1,000 molecular descriptors (3D, 2D, and 1D) for each input structure [108].

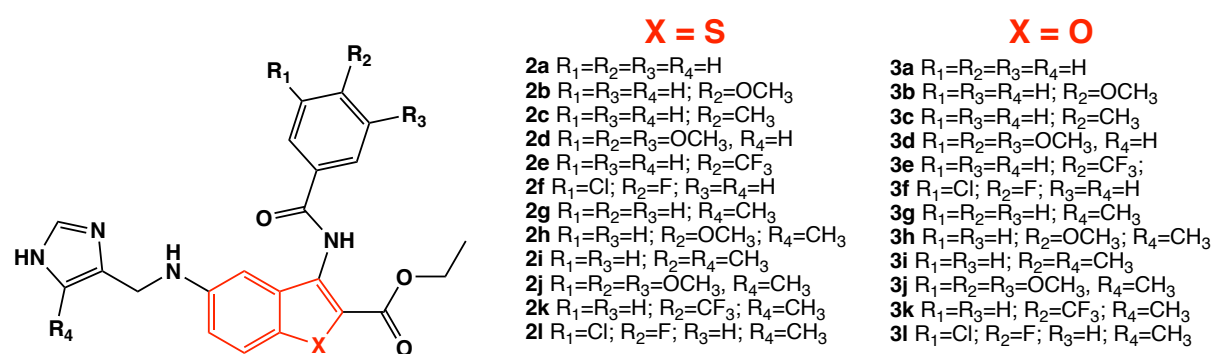
In the context of my PhD thesis, the BPT has been extensively used. It is designed to predict the binding affinity of given input structures against chosen biological targets, available in DRUDIT thanks to a preliminary template building phase. Specifically, sets of known biological target modulators are processed by MOLDESTO, and the full set of molecular descriptors is calculated for each modulator of a defined target. Next, from the output matrices (modulators versus molecular descriptors) mean and standard deviation values for each molecular descriptor are computed. The pairs of values define the biological target template. Independently, when the input structures are submitted to MOLDESTO, the same set of molecular descriptors is computed and matched to the template of the biological target. The

DRUDIT Affinity Score (DAS, which ranges from 0 to 1 indicating low to high affinity) of the input structures versus the biological targets is then computed according to three input parameters N, Z and G which, respectively, control the number of dynamically selected molecular descriptors (N), the maximum allowed percentage of unavailable values (zeros) per molecular descriptor (Z), and the Gaussian smoothing function (G).

The DRUDIT BPT application involved constructing a template for the SARS-CoV-2 M<sup>PRO</sup> binding site, following the described method [9].

Subsequently, an in-house structure database comprising approximately 10,000 heterocyclic structures was uploaded to the DRUDIT PBT. Standard parameters (N=500, Z=50, G=a) were applied, and the output data were ranked according to the DAS, which reflects the compound ability to bind to the SARS-CoV-2 M<sup>PRO</sup> catalytic site.

By applying a cut-off value of 0.8 to DAS, we identified ethyl 3-benzoylamino-5-[(1H-imidazol-4-yl-methyl)-amino]-benzo[*b*]thiophene-2-carboxylate and ethyl 3-benzoylamino-5-[(1H-imidazol-4-yl-methyl)-amino]-benzo[*b*]furan-2-carboxylate, designated as compounds **2** and **3** [109,110] (Figure 12), as promising heterocyclic small molecules for the inhibition of SARS-CoV-2 M<sup>PRO</sup> via modulation of the catalytic active site. Table 1 presents the DAS values of the 24 selected molecules, all characterized by a central heterocyclic benzo[*b*]thiophene or benzo[*b*]furan core with two side moieties: the substituted 3-benzoylamino and the imidazole groups.



**Figure 12.** General structure of ethyl 3-benzoylamino-5-[(1H-imidazol-4-yl-methyl)-amino]-benzo[*b*]thiophene-2-carboxylates **2a-l** and ethyl 3-benzoylamino-5-[(1H-imidazol-4-yl-methyl)-amino]-benzo[*b*]furan-2-carboxylates **3a-l** as new potential antiviral molecules [109,110].

**Table 1.** Biotarget DAS for benzo[*b*]thiophene and benzo[*b*]furan compounds **2a-l** and **3a-l**.

Compound	X	R <sub>1</sub>	R <sub>2</sub>	R <sub>3</sub>	R <sub>4</sub>	DAS
3b	O	H	OCH <sub>3</sub>	H	H	0.940
3a	O	H	H	H	H	0.922
3c	O	H	CH <sub>3</sub>	H	H	0.922
2b	S	H	OCH <sub>3</sub>	H	H	0.918
3g	O	H	H	H	CH <sub>3</sub>	0.916
3f	O	Cl	F	H	H	0.910
3l	O	Cl	F	H	CH <sub>3</sub>	0.910
3i	O	H	CH <sub>3</sub>	H	CH <sub>3</sub>	0.904
3h	O	H	OCH <sub>3</sub>	H	CH <sub>3</sub>	0.900
2d	S	OCH <sub>3</sub>	OCH <sub>3</sub>	OCH <sub>3</sub>	H	0.900
2c	S	H	CH <sub>3</sub>	H	H	0.898
2f	S	Cl	F	H	H	0.890
2g	S	H	H	H	CH <sub>3</sub>	0.888
2i	S	H	CH <sub>3</sub>	H	CH <sub>3</sub>	0.882
2j	S	OCH <sub>3</sub>	OCH <sub>3</sub>	OCH <sub>3</sub>	CH <sub>3</sub>	0.882
2l	S	Cl	F	H	CH <sub>3</sub>	0.880
3d	O	OCH <sub>3</sub>	OCH <sub>3</sub>	OCH <sub>3</sub>	H	0.880
2a	S	H	H	H	H	0.880
2h	S	H	OCH <sub>3</sub>	H	CH <sub>3</sub>	0.880
3k	O	H	CF <sub>3</sub>	H	CH <sub>3</sub>	0.820
2e	S	H	CF <sub>3</sub>	H	H	0.820
3j	O	OCH <sub>3</sub>	OCH <sub>3</sub>	OCH <sub>3</sub>	CH <sub>3</sub>	0.862
3e	O	H	CF <sub>3</sub>	H	H	0.836
2k	S	H	CF <sub>3</sub>	H	CH <sub>3</sub>	0.800

### 3.1.1.2. ADME properties

The 24 selected molecules were submitted to the SwissADME web-tools (<http://www.swissadme.ch>) [111] considering a set of well consolidated parameters for searching bioactive compounds, such as PAINS filters [112], Lipinski's rules [113], Veber [114], and Egan filters [115]. The analysis of the data, showed in Table 2, highlighted that the benzo[*b*]thiophene and benzo[*b*]furan compounds generally met the expectations in terms of bioactivity. Thirteen of the twenty-four structures have no violations, and all compounds have no PAINS. In light of these considerations, no compounds were excluded for the *in silico* structure-based analysis.

**Table 2.** Drug-likeness parameters calculated for the selected 24 compounds.

Compound	Lipinski #violations	Ghose #violations	Veber #violations	Egan #violations	PAINS #alerts	Total
2a	0	0	0	0	0	0

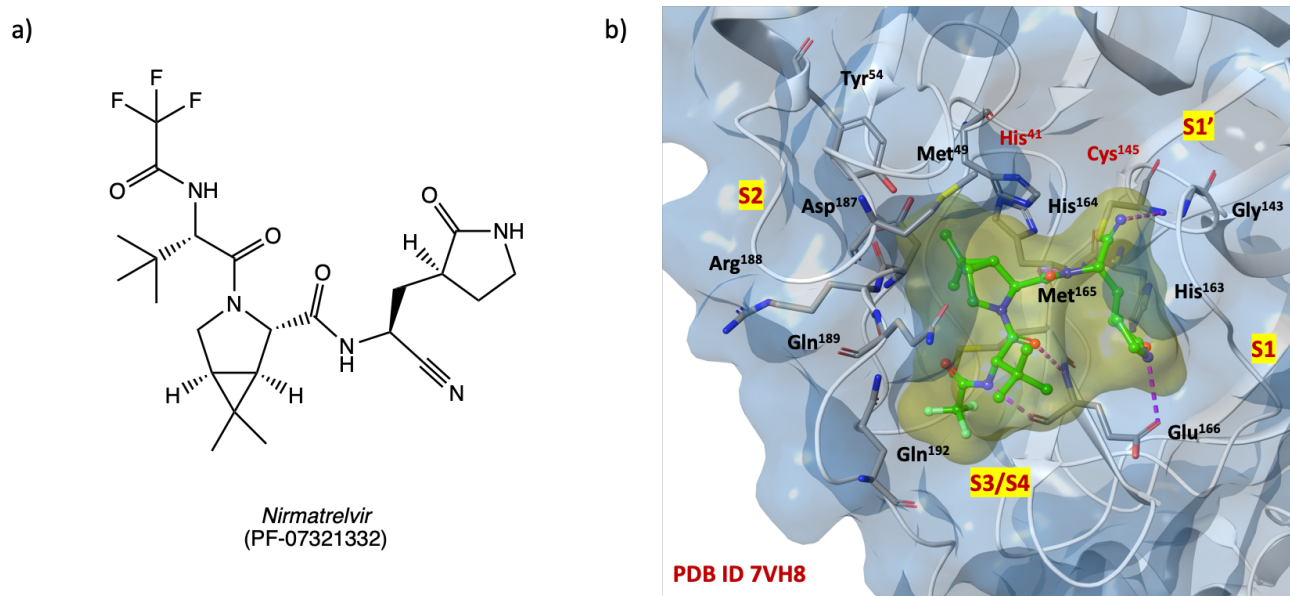
2b	0	0	0	1	0	1
2c	0	0	0	0	0	0
2d	1	2	2	1	0	6
2e	0	2	0	1	0	3
2f	0	0	0	0	0	0
2g	0	0	0	0	0	0
2h	0	0	0	1	0	1
2i	0	0	0	0	0	0
2j	1	2	2	1	0	6
2k	1	2	0	1	0	4
2l	0	2	0	0	0	2
3a	0	0	0	0	0	0
3b	0	0	0	0	0	0
3c	0	0	0	0	0	0
3d	1	2	1	1	0	5
3e	0	1	0	0	0	1
3f	0	0	0	0	0	0
3g	0	0	0	0	0	0
3h	0	0	0	0	0	0
3i	0	0	0	0	0	0
3j	2	2	1	1	0	6
3k	0	2	0	1	0	3
3l	0	0	0	0	0	0

All the parameters calculated through SwissADME are present in the Table S1, Supplementary Material S1.

### 3.1.1.3. *In silico* structure-based studies: Molecular Docking at the catalytic site of SARS-CoV-2 M<sup>PRO</sup>

Induced Fit Docking (IFD) studies were performed to validate the obtained ligand-based data and to gain insight into the structural features of ligand/SARS-CoV-2 M<sup>PRO</sup> (PDB code 7VH8 [116]) complexes, analyzing the mutual conformational changes between ligands and proteins. We focused the docking grid on the SARS-CoV-2 M<sup>PRO</sup> binding pocket, including the four subsites S1', S1, S2, S3/S4 as described in the Materials and Methods section. Figure 13b shows the 3D active binding site of SARS-CoV-2 M<sup>PRO</sup> in covalently bonding with *nirmatrelvir* (PF-07321332, 2D structure in Figure 13a), a second-generation orally available protease

inhibitor currently in phase 3 clinical trials in combination with *ritonavir* (PAXLOVID®, see ClinicalTrials.gov identifier: NCT04960202).



**Figure 13.** (a) 2D structure of *nirmatrelvir*; (b) SARS-CoV-2 M<sup>PRO</sup> 3D binding site surface in complex with PF-07321332 (PDB code 7VH8) [116].

The IFD studies aim to confirm the DRUDIT prediction and the capability of ethyl 3-benzoylamino-5-[(1H-imidazol-4-yl-methyl)-amino]-benzo[*b*]thiophene-2-carboxylates of type **2** and ethyl 3-benzoylamino-5-[(1H-imidazol-4-yl-methyl)-amino]-benzo[*b*]furan-2-carboxylates of type **3** to effectively interact with the selected target binding site. Table 3 shows the IFD and docking scores of the selected 24 structures **2a-l**, **3a-l** and the reference ligand *nirmatrelvir* (PF-07321332) [116].

**Table 3.** IFD and docking output results for **2a-l** and **3a-l** and the co-crystallized ligand *nirmatrelvir* (PDB code 7VH8) [116].

SARS-CoV-2 M <sup>PRO</sup> (PDB code 7VH8)		
Compound	IFD Score	Docking score
<b>2d</b>	-675.768	-8.979
<b>3l</b>	-675.108	-12.040
<b>2j</b>	-674.838	-7.595
<b>2f</b>	-674.292	-9.781
<b>2i</b>	-674.180	-9.222
<b>3i</b>	-674.046	-11.050
<b>3h</b>	-674.040	-10.969
<b>2l</b>	-674.037	-7.673

<b>2a</b>	-673.969	-8.573
<b>2k</b>	-673.927	-10.744
<b>2c</b>	-673.740	-8.008
<b>2b</b>	-673.730	-8.314
<i>nirmatrelvir</i>	-673.142	-10.169
<b>3c</b>	-673.071	-10.733
<b>2h</b>	-673.014	-7.862
<b>3j</b>	-672.880	-9.567
<b>3a</b>	-672.879	-10.861
<b>2g</b>	-672.752	-8.145
<b>2e</b>	-672.547	-8.150
<b>3k</b>	-672.538	-10.328
<b>3g</b>	-672.284	-10.226
<b>3d</b>	-672.184	-9.634
<b>3b</b>	-671.756	-10.415
<b>3f</b>	-671.736	-10.352
<b>3e</b>	-671.460	-9.900

The IFD analysis confirmed the biotarget affinity results and identified the benzo[*b*]thiophenes **2a-d, f, i-l** and benzo[*b*]furans **3h, i, l** as the most promising competitive inhibitors (IFD score range from -675.768 to -673.730) of the SARS-CoV-2 M<sup>PRO</sup> catalytic binding site, with higher IFD scores than *nirmatrelvir* (IFD score -673.142) (Table 3).

Table 4 provides the overview of the amino acids involved in the binding with the 12 highest scoring compounds. The labelled residues were highlighted by the analysis of 2D and 3D ligand pose maps at a distance of 3 Å.

Most of the benzo[*b*]thiophenes **2a-d, f, i-l** and benzo[*b*]furans **3h, i, l** interacts with amino acids Thr<sup>25</sup>, Thr<sup>26</sup>, Leu<sup>27</sup>, Met<sup>49</sup>, Phe<sup>140</sup>, Leu<sup>141</sup>, Asn<sup>142</sup>, Gly<sup>143</sup>, Cys<sup>145</sup>, His<sup>163</sup>, Met<sup>165</sup>, Glu<sup>166</sup>, Pro<sup>168</sup>, Arg<sup>188</sup>, Gln<sup>189</sup>, Thr<sup>190</sup>, Gln<sup>192</sup> with the same reversible interactions observed for the reference compound *nirmatrelvir*. In addition, compounds **2a-d, f, i, k, l** and **3h, i, l** establish more interactions than *nirmatrelvir* with the conserved amino acids of the four SARS-CoV-2 M<sup>PRO</sup> sub-regions S1, S1', S2, and S3/S4, suggesting improved affinity of the benzo[*b*]thiophene and benzo[*b*]furan compounds for the catalytic binding site and resulting more stable ligand/protein complexes.

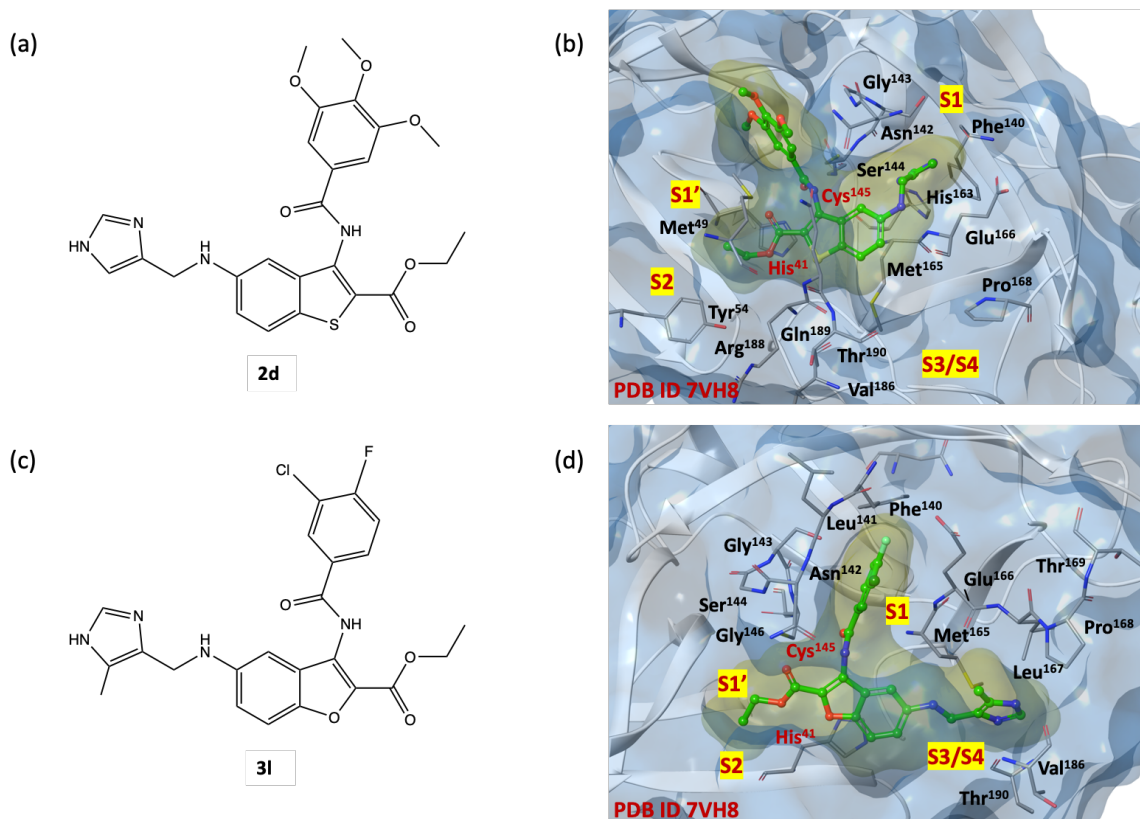
**Table 4.** Overview of the amino acids involved in the binding of the selected 12 compounds with IFD scores higher than *nirmatrelvir* at the SARS-CoV-2 M<sup>PRO</sup> catalytic binding site in proximity of 3 Å.

Title	<b>2d</b>	<b>3l</b>	<b>2j</b>	<b>2f</b>	<b>2i</b>	<b>3i</b>	<b>3h</b>	<b>2l</b>	<b>2a</b>	<b>2k</b>	<b>2c</b>	<b>2b</b>	<i>nirmatrelvir</i>
<b>T25</b>	X	X	X	X	X	X	X	X	X	X	X	X	

<b>T26</b>	X		X	X	X	X	X	X	X	X	X		
<b>L27</b>	X	X	X	X	X	X	X		X	X	X	X	
<b>H41</b>	X	X	X	X		X	X			X	X		
<b>V42</b>	X	X		X									X
<b>C145</b>	X	X	X	X	X	X	X	X	X	X	X	X	X
<b>F140</b>	X	X		X	X	X	X	X	X	X	X	X	
<b>L141</b>	X	X			X	X	X	X	X	X	X		
<b>N142</b>	X	X	X	X	X	X	X	X	X	X	X	X	X
<b>G143</b>			X		X	X	X	X	X	X	X	X	X
<b>H163</b>	X	X		X	X	X		X	X		X	X	X
<b>E166</b>	XX	X	XX	X	X	X	X	X	X	X	X	X	X
<b>H172</b>					X			X	X		X	X	
<b>M49</b>			X	X	X	X	X	X	X		X	X	X
<b>M165</b>	X	X	X	X	X	X	X	X	X	X	X	X	X
<b>L167</b>		X			X	X	X	X		X			X
<b>P168</b>		X			X	X	X	X	X	X	X	X	X
<b>V186</b>		X			X	X							
<b>D187</b>	X	X	X	X									X
<b>R188</b>	X	X	X	X	X	X	X	X	X	X	X	X	X
<b>Q189</b>	X	X	X	X	XX	X	X	X	X	X	X	X	X
<b>T190</b>		X			X	X	X	X	X	X	X	X	X
<b>Q192</b>		X			X	X	X	X		X		X	X
<b>TOT</b>	16	19	14	16	21	20	18	18	17	17	18	17	15

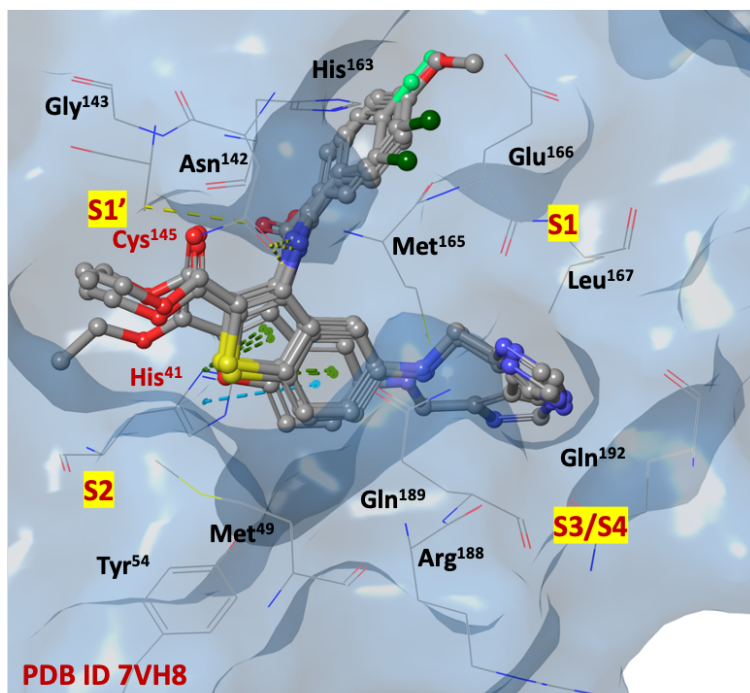
As shown in Figures 14b,d, the binding pocket exhibits suitable properties for ligands **2d** and **3l**, which are the two highest IFD scoring compounds, (2D structures shown in Figures 14a,c). Both compounds interact with the SARS-CoV-2 M<sup>PRO</sup> binding site by extending all substituents into the four subsites S1', S1, S2, S3/S4, and creating a network of key reversible hydrogen bonds with the amino acids: His<sup>41</sup>, Phe<sup>140</sup>, Asn<sup>142</sup>, Cys<sup>145</sup>, Glu<sup>166</sup>, Gln<sup>189</sup>, Thr<sup>190</sup>, Tyr<sup>54</sup>, Met<sup>165</sup>, Val<sup>186</sup> and His<sup>41</sup>, Phe<sup>140</sup>, Asn<sup>142</sup>, Cys<sup>145</sup>, Glu<sup>166</sup>, Gln<sup>189</sup>Thr<sup>190</sup>, Tyr<sup>54</sup>, Met<sup>165</sup>, Val<sup>186</sup> for benzo[*b*]thiophene **2d** and benzo[*b*]furan **3l**, respectively.





**Figure 14.** (a) 2D structure of benzo[*b*]thiophene **2d**; (b) 3D complex of SARS-CoV-2 M<sup>PRO</sup> binding site (PDB code 7VH8) with benzo[*b*]thiophene **2d** [116]; (c) 2D structure of benzo[*b*]furan **3l**; (d) 3D complex of SARS-CoV-2 M<sup>PRO</sup> binding site (PDB code 7VH8) with benzo[*b*]furan **3l** [116].

Among the ligands with higher IFD scores, the analysis of the binding poses of derivatives **2b**, **2c**, **2l** and **3l**, Figure 15, highlights a remarkable overlap of poses, indicating a redundancy in the position of the key elements of the small molecules within the four sub-pockets, and forming many interactions. The imidazole moiety, charged at physiological pH, is capable to penetrate deeply into the S1 and S3/S4 sub-regions (Asn<sup>142</sup>, Gly<sup>143</sup>, His<sup>163</sup>, His<sup>164</sup>, Glu<sup>166</sup>) and stabilize itself by forming hydrogen interactions with the side chains and/or the backbone of residues, that are particularly exposed to the solvent in the S3/S4 site. Probably, this portion could mimic the Gln residue of the natural substrates, similarly to the five membered  $\gamma$ -lactamic ring, which is the most recurrent fragment of selective inhibitors at the catalytic binding site, reported in the literature [76].



**Figure 15.** 3D overlaps of compounds **2b**, **2c**, **2l** and **3l** at SARS-CoV-2 M<sup>PRO</sup> binding site (pdb code 7VH8) [116].

The SARS-CoV-2 M<sup>PRO</sup> S2 cleft, comprised mainly of hydrophobic amino acids (Met<sup>49</sup>, Tyr<sup>54</sup>, Met<sup>165</sup>, Leu<sup>167</sup>, Pro<sup>168</sup>, Val<sup>186</sup>, Asp<sup>187</sup>, Arg<sup>188</sup>), appears to be very flexible, allowing it to bind both small and bulky aromatic/alkyl portions. The heterocyclic scaffold could represent a central pharmacophoric portion, thanks to its capability to stabilize the ligand/protein complex in favorable ligand positions. Additional  $\pi$ - $\pi$  stacking interactions are observed between the benzo[*b*]thiophene and benzo[*b*]furan ring systems and the imidazole substituent of His<sup>41</sup>.

The carboxyethyl moiety is stabilized in the S1' pocket, instead the carboxamide moiety, simulating a peptide bond of natural substrate creates several interactions with the side chains of His<sup>41</sup>, Asn<sup>142</sup> and Glu<sup>166</sup>, in both S1' and S1 pockets. The substituted phenyl rings are arranged in the region adjacent to the S1 cleft.

#### 3.1.1.4. Statistical Analysis: Principal Component Analysis

The structural features of the selected compounds of types **2** and **3** prompted us to evaluate them also as potential binders for the allosteric dimerization site of the SARS-CoV-2 M<sup>PRO</sup>, as non-competitive inhibitors. Like the known dimerization site inhibitor *pelitinib*, the

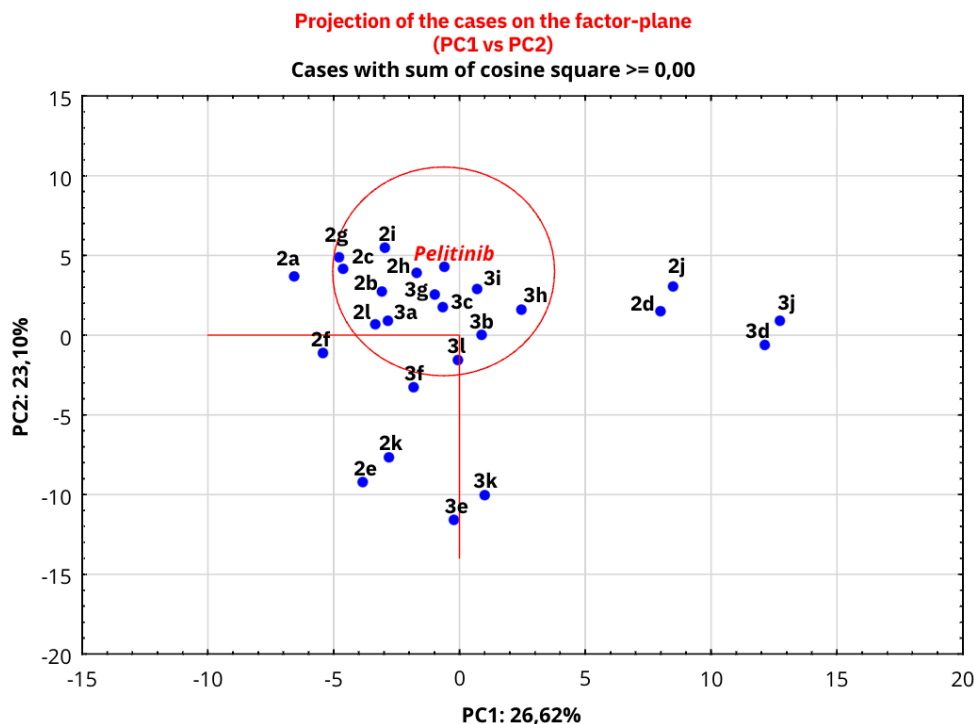
benzo[*b*]thiophene and benzo[*b*]furan compounds have a series of aromatic rings, linked by rotatable bonds.

To investigate the potential inhibitory activity of the selected compounds at the dimerization site of SARS-CoV-2 M<sup>PRO</sup>, we performed a Principal Component Analysis (PCA), including the molecular descriptor matrix of the 24 compounds, obtained from DRUDIT ligand-based studies, and the molecular descriptors of the non-covalent allosteric inhibitor *pelitinib*.

PCA is a technique used for data simplification in multivariate statistics, aimed at reducing the number of variables that describe a dataset while minimizing information loss. This reduction is achieved by transforming the original matrix into a new one, where the variables, called Principal Components (PCs), are orthogonal (uncorrelated) and arranged so that the first components, in descending order of importance, retain most of the variance contained in the original set of variables. The number of PCs can be equal to or fewer than the number of variables in the dataset. The first PC is defined as the one that contributes the most. The second PC is the best linear combination of variables that represents the maximum possible residual variance after removing the effect of the first component, and so on.

In a 2D plot of PC1 vs. PC2, objects with similar properties can be clustered in specific areas of the Cartesian plane

The application of PCA to the matrix of structures *versus* Molecular Descriptors (Supplementary Material S1, Matrix S1), showed a total variance of 50% expressed by the first two components. The bidimensional plot (PC1 versus PC2, Figure 16) shows the 2D arrangement of the molecules in the graph compared with the reference ligand *pelitinib*.



**Figure 16.** Principal Component Analysis (PC1 versus PC2) applied to the molecular descriptors matrix of the selected compounds and *pelitinib*.

Based on the distribution of the compounds in this graph, the molecules in the proximity of *pelitinib* were identified as new inhibitors potentially capable of modulating the dimerization process. The distances of the molecules from the reference *pelitinib* were calculated and the data were listed in Table 5, which provides cartesian coordinates for the derivatives in the circle (Figure 16), (see Table S2, Supplementary Material S1, for all coordinates). Benzo[*b*]thiophenes **2b,c,g-i,l** and Benzo[*b*]furans **3a-c,g-i,l** were selected to be further investigated with structure-based studies, considering the dimerization region of SARS-CoV-2 M<sup>PRO</sup> as allosteric binding site.

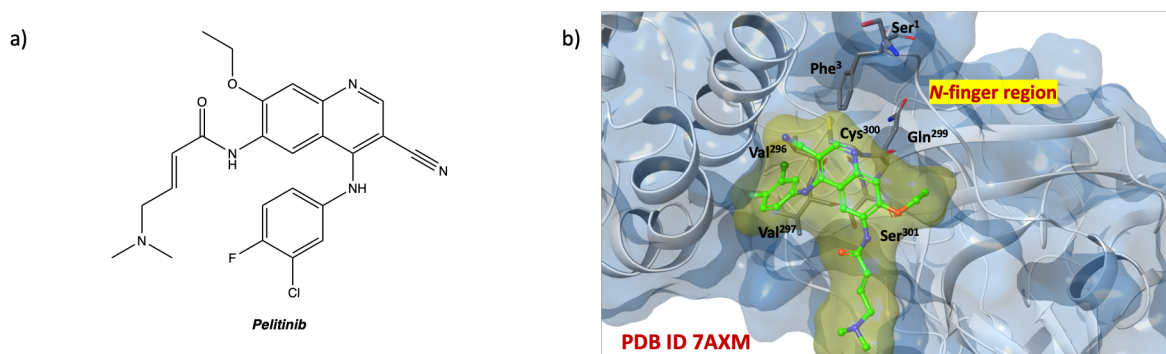
**Table 5.** Distances calculated for each input molecules from *pelitinib*.

Compound	PC1	PC2	Distance
<i>pelitinib</i>	-0.62	4.23	-
<b>2h</b>	-1.70	3.92	1.15
<b>3g</b>	-0.98	2.56	1.76
<b>3i</b>	0.70	2.91	1.90

<b>3c</b>	-0.67	1.76	2.53
<b>2i</b>	-2.98	5.49	2.66
<b>2b</b>	-3.08	2.73	2.92
<b>2c</b>	-4.63	4.16	4.02
<b>3a</b>	-2.85	0.91	4.05
<b>3h</b>	2.46	1.59	4.09
<b>2g</b>	-4.78	4.90	4.22
<b>2l</b>	-3.35	0.69	4.53
<b>3b</b>	0.87	0.01	4.53
<b>3l</b>	-0.06	-1.55	5.86

### 3.1.1.5. *In silico* structure-based studies: Induced Fit Docking into the allosteric site of SARS-CoV-2 M<sup>PRO</sup>

Following to the statistical analysis, IFD simulations, fixing the docking grid on the SARS-CoV-2 M<sup>PRO</sup> allosteric site on PDB code 7AXM [83] were performed. In Figure 17b, the crystallographic structure of SARS-CoV-2 M<sup>PRO</sup> complexed with non-covalent allosteric inhibitor *pelitinib* is shown (2D structure in Figure 17a).



**Figure 17.** (a) 2D structure of *pelitinib*; (b) 3D complex of SARS-CoV-2 M<sup>PRO</sup> allosteric site with *pelitinib* (PDB code 7AXM) [83].

The analysis of the results shows compounds with higher affinity than *pelitinib*. In Table 6 the IFD and docking values of the selected derivatives and of the reference compound are reported. In particular, the benzo[*b*]thiophenes **2b**, **c**, **g**, **i**, **l** and benzo[*b*]furans **3b**, **i**, **l** exhibit

IFD scores in the range from -675.768 to -673.730 suggesting interesting allosteric affinity for the dimerization binding site of the SARS-CoV-2 M<sup>PRO</sup>.

**Table 6.** IFD and docking output results (pdb code 7AXM) [83].

Compound	IFD Score	Docking Score
<b>2c</b>	-693.48	-7.005
<b>2b</b>	-692.66	-7.214
<b>3l</b>	-692.66	-6.327
<b>2l</b>	-692.59	-6.72
<b>3i</b>	-692.24	-7.522
<b>2g</b>	-692.05	-6.368
<b>2i</b>	-691.57	-5.981
<b>3b</b>	-691.36	-6.187
<i>pelitinib</i>	-691.09	-6.192
<b>3c</b>	-691.03	-6.675
<b>2h</b>	-690.98	-5.082
<b>3g</b>	-690.73	-5.954
<b>3h</b>	-690.62	-5.679
<b>3a</b>	-689.92	-6.238

Considering the IFD results for both catalytic and dimerization sites, compounds **2b**, **c**, **i**, **l** and **3i**, **l** were found to have interesting IFD scores, suggesting a dual binding site inhibitory activity of SARS-CoV-2 M<sup>PRO</sup>.

Analysis of the amino acid maps (Table 7), in combination with the 2D and 3D poses examination of the ligands on the dimerization site uncovered promising expectations. Most benzo[*b*]thiophenes **2b**, **c**, **i**, **l** and benzo[*b*]furans **3i**, **l** interact with amino acids Thr<sup>154</sup>, Pro<sup>252</sup>, Gln<sup>256</sup>, Val<sup>297</sup>, Arg<sup>298</sup>, Cys<sup>300</sup>, Val<sup>303</sup>, Thr<sup>304</sup>, with similar reversible interactions observed with the reference compound *pelitinib*.

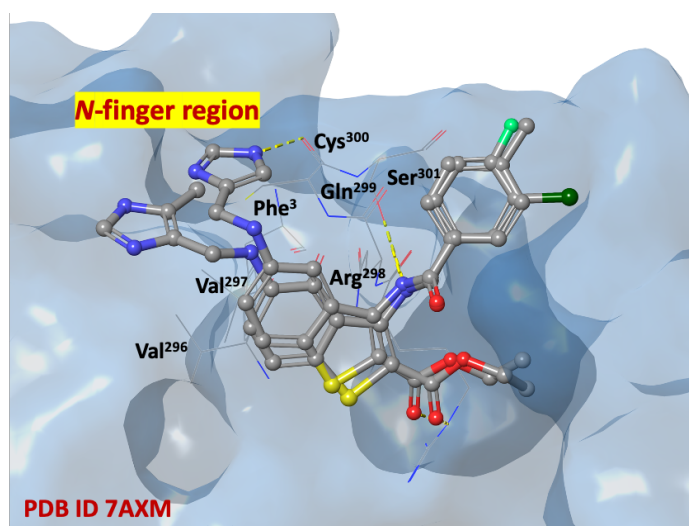
**Table 7.** Overview of the amino acids involved in the binding with the selected 13 compounds at the SARS-CoV-2 M<sup>PRO</sup> allosteric site in proximity of 3 Å.

Title	<b>2c</b>	<b>2b</b>	<b>3l</b>	<b>2l</b>	<b>3i</b>	<b>2g</b>	<b>3h</b>	<b>2i</b>	<b>3b</b>	<i>pelitinib</i>
<b>S1</b>		X								
<b>G2</b>		X								
<b>D153</b>	X			X		X				
<b>Y154</b>	X		X	X		X	X		X	
<b>T209</b>					X			X		
<b>A210</b>					X			X		

<b>I213</b>		X			X			X	X	
<b>N214</b>		X								
<b>I249</b>			X							
<b>P252</b>	X	X	X	X	X	X			X	X
<b>L253</b>		X			X			X		X
<b>A255</b>				X	X				X	
<b>Q256</b>	X	X	X	X	X	X		X	X	X
<b>F294</b>			X			X	X			
<b>V296</b>					X			X		X
<b>V297</b>	X	X	X	X	X	X	X	X	X	X
<b>R298</b>	X		X	X		X	X		X	
<b>C300</b>	X	XX		X	X			X	X	X
<b>S301</b>	X	X	XX		XX	X	X	X	X	X
<b>G302</b>			X		X		X	X	X	
<b>V303</b>	X	X	X	X	X	X	X	X	X	X
<b>T304</b>	X		X	X	X	X	X		X	
<b>F305</b>				X		X				
<b>TOT</b>	10	12	12	11	15	11	8	11	12	8

The selected compounds establish hydrogen bonds with both the residues of the *N*-finger region and the region around residues Arg<sup>298</sup>, Cys<sup>300</sup>, Ser<sup>301</sup>. The allosteric ligand/protein complex in the pocket between the domains II and III could interfere with the necessary interactions between the two monomers and prevent the dimerization process with the consequent inactivation of the SARS-CoV-2 M<sup>PRO</sup>.

In Figure 18 are shown compounds **2c** and **2l**, in a recurring position, fitting in the SARS-CoV-2 M<sup>PRO</sup> allosteric site (dimerization domain), in which there are highly conserved and essential amino acids, such as *N*-finger residues (Ser<sup>1</sup>, Gly<sup>2</sup>, Phe<sup>3</sup>, Arg<sup>4</sup>, Lys<sup>5</sup>, Met<sup>6</sup>, Ala<sup>7</sup>), Pro<sup>293</sup>, Phe<sup>294</sup>, Asp<sup>295</sup>, Val<sup>296</sup>, Val<sup>297</sup>, Arg<sup>298</sup>, Gln<sup>299</sup>, Cys<sup>300</sup>, Ser<sup>301</sup>.



**Figure 18.** 3D overlaps of compounds **2c**, and **2l** at the SARS-CoV-2 M<sup>PRO</sup> allosteric site (PDB code 7AXM) [83].

### 3.1.1.6. Enzyme inhibition assay of SARS-CoV-2 M<sup>PRO</sup>

The activity of M<sup>PRO</sup> was assessed using the SensoLyte 520 SARS-CoV-2 3CL<sup>PRO</sup> Activity Assay Kit. Compounds were serially diluted from 5000  $\mu\text{M}$  to 1  $\mu\text{M}$  and tested for their ability to inhibit the protease activity. The assay results revealed that compound **2i-l** exhibited IC<sub>50</sub> value lower of 100  $\mu\text{M}$  when testing on the SARS-CoV-2 M<sup>PRO</sup> catalytic site. In details, IC<sub>50</sub> values for **2i-l** are of 70.4  $\mu\text{M}$ , 87.1  $\mu\text{M}$ , 60.8  $\mu\text{M}$ , and 45.9  $\mu\text{M}$ , respectively. This, as suggested from our *in silico* analysis in section 3.1.1.3., resulted as agents with potential inhibitory effect. Moreover, among them, compounds **2i** (IC<sub>50</sub> values of 70.4) and **2l** (IC<sub>50</sub> values of 45.9) also showed favorable activity as dual binding inhibitors, positioning them as promising lead compounds for further development as antiviral agents.

In conclusions, compounds **2b,c,i,l** and **3i,l**, with favorable ADME properties (drug-likeness, lead-likeness, no PAINS) and already known antitumoral activity [109,110], showed capability to strongly bind both to the catalytic and allosteric SARS-CoV-2 M<sup>PRO</sup> sites suggesting a potential dual activity. In general, considering the high amino acids conservation rate in both pockets, the potential drugs proposed here might even be effective against mutation variants and other coronaviruses. This analysis improves our knowledge of protein–ligand relationships in SARS-CoV-2 M<sup>PRO</sup> catalytic and allosteric sites and offer new molecular scaffolds for inhibitor design.



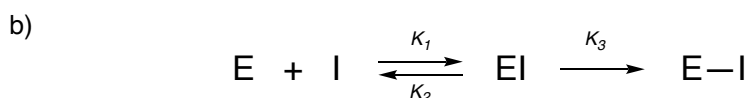
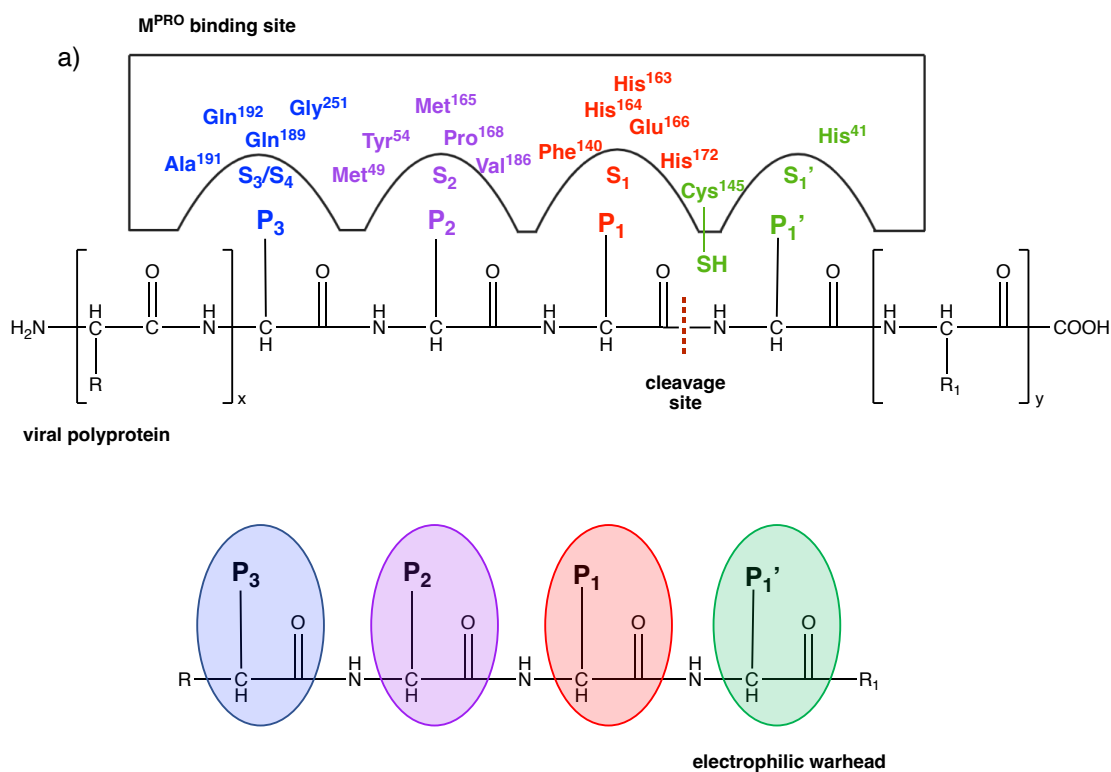
## 3.2. Targeting SARS-CoV-2 M<sup>PRO</sup> through Covalent Inhibition

### 3.2.1. Targeting Sars-Cov-2 Main Protease for the Treatment of COVID-19: Covalent inhibitors Structure-Activity Relationships insights and Evolution Perspectives

Covalent inhibition strategies have previously shown significant success against other viral proteases, including those for MERS and SARS-CoV-1 [117]. Unlike traditional non-covalent inhibitors, covalent agents often offer enhanced efficacy, greater potency, prolonged residence time in the binding site, sustained pharmacological action, and the potential to overcome resistance [118-120]. In the context of SARS-CoV-2 M<sup>PRO</sup>, the catalytic cysteine residue (Cys<sup>145</sup>) presents a prime target for covalent inhibition. Electrophilic warhead inhibitors, designed to react with this catalytic site, mimic the amide peptide bonds of the viral polyproteins, thereby facilitating their binding and subsequent inactivation of the enzyme.

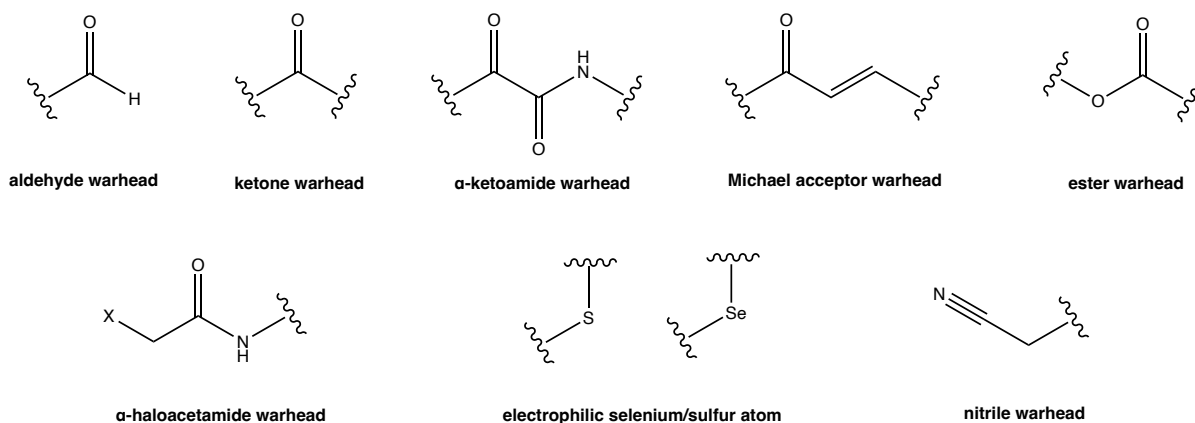
For covalent inhibition to be successful, inhibitors must meet basic structural requirements: a P1 portion, usually a cyclic glutamine analogue, capable of interacting with amino acids such as His<sup>163</sup>, Glu<sup>166</sup>, His<sup>172</sup> via hydrogen bonds and hydrophobic interactions; P2 and P3 moieties, projected into hydrophobic pockets (S2 and S3/S4) and characterized by substituent groups capable of forming favorable hydrophobic interactions with the amino acids of these sites, (Figure 19a).

In this perspective, the kinetic scheme of covalent inhibition deserves special attention. As illustrated in Figure 19b, kinetic inhibition involves two steps: firstly, the inhibitor (I) reversibly binds into the active site of the enzyme (E), forming a non-covalent enzyme-inhibitor complex (EI); then, following the nucleophilic attack by the catalytic Cys<sup>145</sup>, the non-covalent complex evolves in a stable complex (E—I). The first stage is governed by the equilibrium-binding constant  $K_i$  ( $K_i = k_2/k_1$ ), while the second one by the inactivation constant for covalent bond formation  $k_3$  ( $K_{inact}$ ) [121].



**Figure 19. (a)** Covalent inhibition of the SARS-CoV-2 M<sup>PRO</sup>. The P1'-P1-P2-P3 labels reflect the chemical analogies with the viral substrate. The warheads P1' in green, while the fragments P1-P3 are shown in red, purple, and blue, respectively. The sub-regions of the binding pocket are labeled with S numbering complementary to the fragments of the inhibitor; **(b)** kinetic scheme of covalent inhibition. E, I, EI and E-I, represent enzyme, Inhibitor, enzyme-inhibitor non-covalent complex and enzyme-inhibitor covalent complex, respectively.

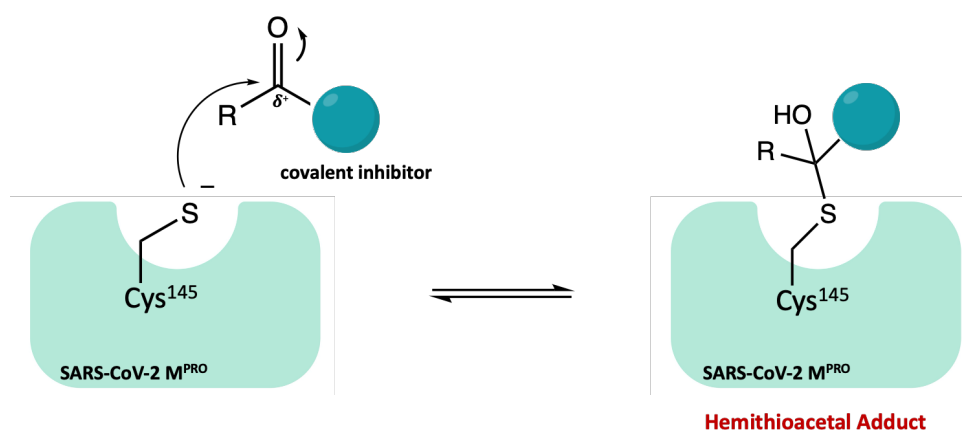
Over the past four years, substantial efforts have been dedicated to developing new covalent inhibitors targeting SARS-CoV-2 M<sup>PRO</sup> [76]. Several small molecules have advanced to the more critical stages of drug development, reflecting their potential in treating COVID-19. These covalent inhibitors are categorized based on the nature of their electrophilic warheads into eight groups: aldehydes, ketones,  $\alpha$ -ketoamides, Michael acceptors,  $\alpha$ -haloacetamides, nitriles, esters, and molecules containing electrophilic selenium/sulfur atoms (Figure 20). Most covalent inhibitors exhibit a dipeptidomimetic or tripeptidomimetic structure. However, some non-peptidomimetic SARS-CoV-2 M<sup>PRO</sup> inhibitors are also reported in the literature, including activated ester derivatives, natural compounds, or *eb sulfur/eb selen* derivatives.



**Figure 20.** Electrophilic warheads characterizing covalent SARS-CoV-2 M<sup>PRO</sup> inhibitors.

### 3.2.1.1. Carbonyl warhead

Carbonyl groups, such as aldehydes and ketones, are considered the most promising warheads in the design of new covalent inhibitors of SARS-CoV-2 M<sup>PRO</sup>. From a mechanistic point of view, their capability to form a covalent bond depends on the electrophilicity of the carbonyl carbon, which is susceptible to the nucleophilic addition of the cysteine-SH, leading to the formation of a reversible hemithioacetal adduct. The high similarity between the latter and the intermediate formed by the natural substrate during the enzymatic catalytic cycle ensures high stability of the inhibitor-protein complex and longer residence time [72]. Figure 21 shows the general mechanism of covalent inhibition for this class of compounds.



**Figure 21.** General mechanism of action of covalent carbonyl inhibitors.

#### 3.2.1.1.1. Aldehyde warhead

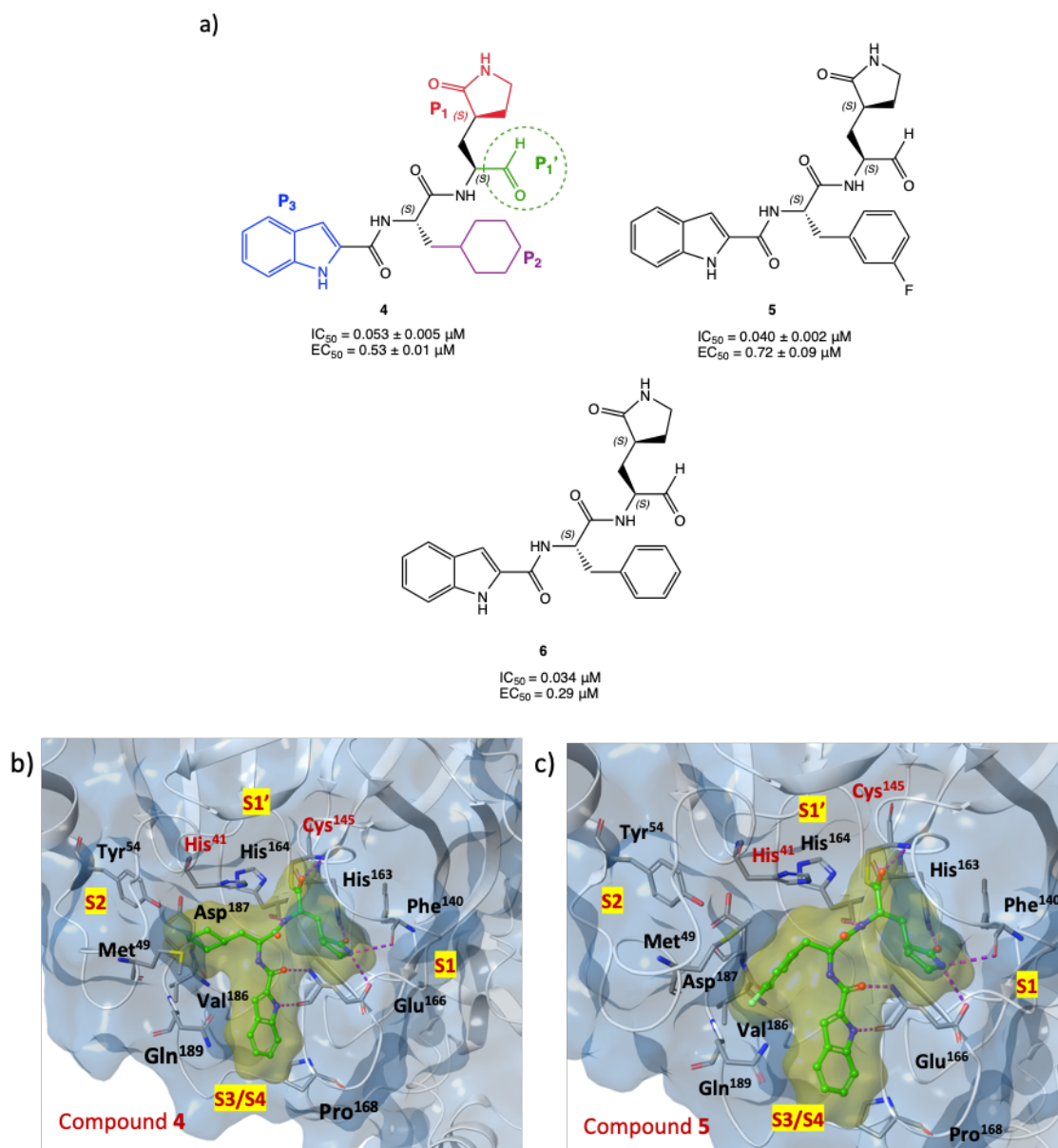
Compounds **4** and **5**, in Figure 22a, are among the first peptidomimetic covalent inhibitors with an aldehyde group as a warhead. *In vitro* inhibition assays proved excellent inhibitory

activity against SARS-CoV-2 M<sup>PRO</sup> (IC<sub>50</sub> of 0.053 ± 0.005 μM and 0.040 ± 0.002 μM, respectively), strong anti-SARS-CoV-2 infectivity (EC<sub>50</sub> of 0.53 ± 0.01 μM and 0.72 ± 0.09 μM, respectively), low cytotoxicity and favorable pharmacokinetic, and toxicity properties *in vivo*. Interestingly, the X-ray structures of both compounds in complex with M<sup>PRO</sup>, resolved at 1.5 Å, allowed a detailed analysis of the binding mode and mechanism of action of the proposed leads. In Figure 22b, c the SARS-CoV-2 M<sup>PRO</sup> in complex with **4** and **5** (PDB code: 6LZE and PDB code: 6MOK, respectively), is shown. To graphically illustrate the importance of each ligand moiety in stabilizing the complex, the four binding site cavities and the corresponding amino acids involved in the key interactions were highlighted. From the analysis, a similar binding mode of the two ligands was found: in both complexes, the formation of the covalent bond between the aldehyde group (P1') and the Cys<sup>145</sup>-SH at S1' is enhanced by the additional H-bond between the hemithioacetal group -OH and the cysteinyl backbone; the (S)-γ-lactam group (P1) is deeply inserted into the S1 site and, mimicking the glutamine residue of the natural substrates, can form three H-bonds with three key amino acids of this cleft (one between His<sup>163</sup> and the lactamic oxygen, two between Phe<sup>140</sup> and Glu<sup>166</sup> and the lactamic NH); the indole group (P3) is located at the surface (S3/S4 pocket) of the protein and interact with Glu<sup>166</sup>. The only observed difference in the binding mode is represented by the P2 fragments, that differ between the two compounds (cyclohexyl and 3-fluorophenyl groups, respectively). Indeed, the 3-fluorophenyl moiety is more down-rotated than the P2 in compound **4**. However, despite this difference in orientation, both fragments are deeply inserted into the S2 cavity and form extensive hydrophobic interactions [122].

Furthermore, in the light of the interesting results, derivative **4** was chosen as a representative compound to gain insight into the mechanism of action of aldehyde inhibitors at atomistic level using QM/MM simulations: this type of molecular simulations permitted to elucidate the mechanism of covalent bond formation and to highlight the proton transfer processes, that take place within the catalytic dyad prior to the nucleophilic attack [123].

Furthermore, searching for new inhibitors of rhinovirus and enterovirus proteases, synthesised derivative **6** (Figure 22a), a peptidomimetic analogue of **5**, without the 3-F substitution on the phenyl ring. Tested against SARS-CoV-2 M<sup>PRO</sup>, compound **6** exhibited interesting inhibitory activity with IC<sub>50</sub> value of 0.034 μM and antiviral activity with an EC<sub>50</sub> = 0.29 μM. Structure-Activity Relationship (SAR) studies confirmed the importance of the γ-

lactamic pentacyclic system and the indole ring: the substitution with other heterocycles, such as quinoline, showed a drastic decrease in activity [124].



**Figure 22.** (a) Chemical structures of peptidomimetic aldehyde derivatives 4-6; for derivative 4, as an example for the whole series, the four key moieties P1', P1, P2, and P3 for the interaction with SARS-CoV-2 M<sup>PRO</sup> are highlighted; (b) X-ray structure of SARS-CoV-2 M<sup>PRO</sup> in complex with compound 4 (PDB code: 6LZE); (c) X-ray structure of SARS-CoV-2 M<sup>PRO</sup> in complex with compound 5 (PDB code: 6MOK). In both crystal structures the amino acids involved in the interaction with the ligand and the four cavities of the binding pocket (S1', S1, S2, S3/S4) are showed [122-124].

Other examples of promising SARS-CoV-2 M<sup>PRO</sup> covalent inhibitors with aldehyde warhead include GC-373 and the related prodrug GC-376, in which the aldehyde function is hidden as bisulphite group to increase the solubility, and which can be readily released under physiological conditions (Figure 23a).

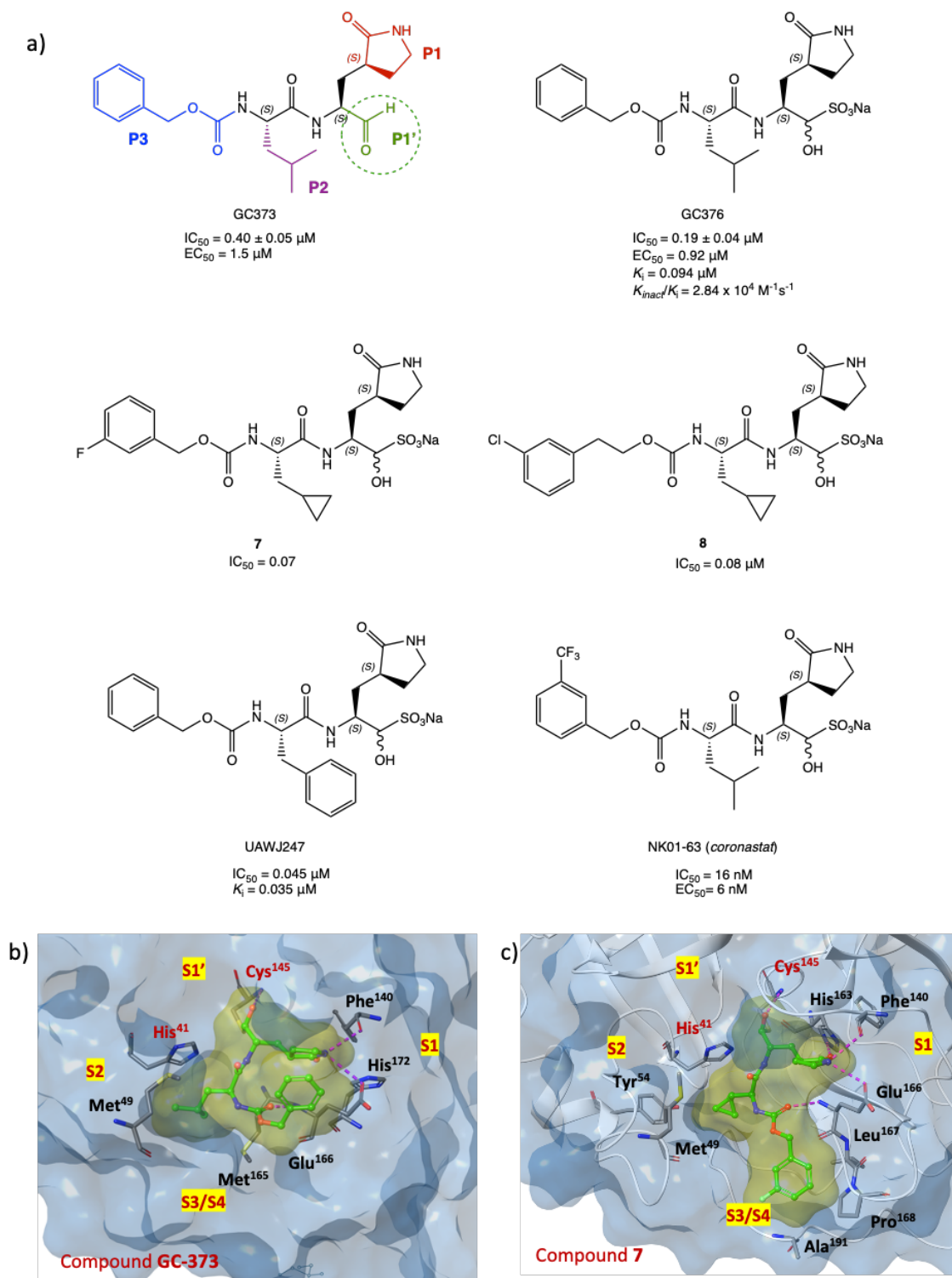
These two compounds, initially investigated as veterinary drugs for their capability to inhibit of Feline CoronaVirus (FCoV) 3CL protease and proposed for the treatment of feline infectious peritonitis [125], were repurposed as new anti-COVID-19 agents, showing potent inhibitory activity on M<sup>PRO</sup>. Specifically, GC-373 exhibited an IC<sub>50</sub> value against SARS-CoV-2 M<sup>PRO</sup> of 0.40 ± 0.05 μM, while the prodrug GC-376 produced slightly higher inhibition, with an IC<sub>50</sub> of 0.19 ± 0.04 μM [126,127]. X-ray studies performed on both compounds in complex with the SARS-CoV-2 M<sup>PRO</sup> showed an identical binding mode of the ligands (for GC-373 PDB code: 6WTK; for GC-376 PDB code: 6WTJ), confirming the ability of the bisulphite adduct of GC-376 to rapidly release the aldehydic function. To gain insight into the mechanism of inhibition and the binding mode of this class of compounds, the crystal structure of GC-376 in complex with the target protein is shown in Figure 23b, highlighting the key structural components responsible for the inhibition: the aldehydic warhead (P1') forms a covalent hemithioacetal adduct with the cysteinyl -SH within the catalytic cleft in S1'; the leucine P2 inserts itself into the hydrophobic subregion S2 of the enzyme; the recurrent γ-lactam P1 of the substrate is located into the S1 site and forms multiple H-bonds with polar amino acids (e.g. His<sup>163</sup> and Glu<sup>166</sup>); the benzyl group in P3 is projected into the superficial S3/S4 site of the protein, where it forms an H-bond with the Glu<sup>166</sup> backbone [126].

Further *in vitro* studies were performed on a model of SARS-CoV-2 infection in Vero-E6 cells, and confirmed the high inhibitory activity of both GC-373 and GC-376, which exhibited an EC<sub>50</sub> of 1.5 μM and 0.92 μM, respectively, and low cytotoxicity even at high concentration [126,127]. The remarkable antiviral effect in the cellular assay (also measured in cells infected with delta and omicron SARS-CoV-2 variants) was associated with the dual-effect of inhibition of the host cysteinyl protease cathepsin L, which is involved in the infection process of SARS-CoV-2 into the host cells along with other proteases such as calpains [128-130].

Inspired by the *in vitro* results and the previous positive outcomes in animals affected by FCoV, Caceres and co-workers examined GC-376 *in vivo* in the K18 hACE2/SARS-CoV-2 transgenic mouse model, which expresses the human angiotensin-converting enzyme type 2 and is affected by SARS-CoV-2. The analysis of the data showed modest benefit in terms of clinical symptoms and survival, but an interesting capability to reduce tissue lesions and inflammation in the tested animals [131].

With the aim of improving the *in vitro* inhibition effect of GC-376 against M<sup>PRO</sup>, a series of analogues have been designed. From the viewpoint of SAR, the structural modifications have

been directed mainly towards the P2 (leucine isopropyl) and the P3 (carboxy-benzyl) moieties; the P1 portion ( $\gamma$ -lactam) and the P1' aldehyde warhead were retained in view of their pivotal role in the interaction with the target. The most representative compounds were the bisulphite prodrugs **7** and **8** (Figure 23a), which were modified at both P2 and P3 sites (cyclopropyl group and halogen-substituted phenyl, respectively). *In vitro* inhibition assay proved their better inhibitory activity against M<sup>PRO</sup> (IC<sub>50</sub> of 0.07 and 0.08  $\mu$ M, respectively) and their antiviral effects (EC<sub>50</sub> of 0.57 and 0.7  $\mu$ M, respectively) compared to both the lead compound GC-376 and the corresponding parent aldehydes. The interesting potency of compound **7** was confirmed by the analysis of the X-ray complex of M<sup>PRO</sup> with the desalted form of **7**. As shown in Figure 23c (complex of derivative **7** with SARS CoV-2 M<sup>PRO</sup>, PDB code: 7LCO), the introduction of the new structural features contributes to enhance the interaction with the target protein and achieve a better fit into the active site compared to GC-376. The more compact cyclopropyl group on the P2 unit of the ligand is able to penetrate deeper into the S2 pocket of the target; furthermore, the introduction of a halogen substituent on the Cbz group (P3 fragment of the ligand) permits to move this moiety closer to the S3/S4 cleft, while in the GC-373 (Figure 23b), the unsubstituted Cbz is mainly directed towards the surface of the protease exposed the solvent [132].



**Figure 23. (a)** Chemical structures of GC-373 (electrophilic P1' warheads in green, P1 in red, P2 in purple, P3, in blue), GC-376, analogues **7**, **8**, UAWJ247, and NK01-63 (*coronastat*) [125,132,133]; **(b)** X-ray structure of SARS-CoV-2 M<sup>PRO</sup> in complex with GC-373 (PDB code: 6WTJ [126]); **(c)** X-ray structure of SARS-CoV-2 M<sup>PRO</sup> in complex with **7** (PDB code: 7LCO [132]).

To further, explore the structural determinants of GC-376 for more selective inhibition of SARS-CoV-2 M<sup>PRO</sup>, compound UAWJ247 (Figure 23a) was designed by replacing the isopropyl



moiety with a phenyl group. The *in vitro* inhibition assays registered IC<sub>50</sub> value very close to that of the parent compound (0.045 μM), confirming a better fitting of the bulkier aromatic portion into the S2 site of the M<sup>PRO</sup> binding pocket (crystal structure complex with PDB code: 6XBH) [133].

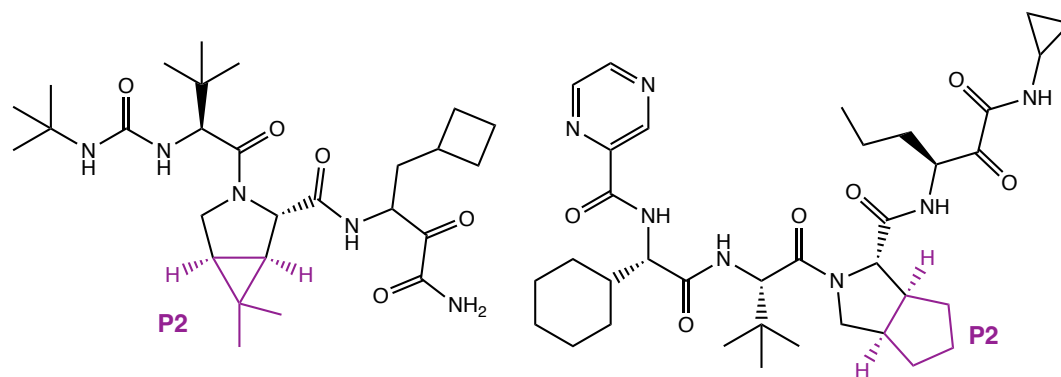
Recently, in search of new GC-376 analogues with improved inhibitory and drug-like properties, a series of compounds modified at the P2 and P3 sites was designed. Among them, the compound NK01-63 (*coronastat*), proved to be more active than the lead compound. It showed excellent inhibitory activity against SARS-CoV-2 (IC<sub>50</sub> = 16 nM), potent antiviral effect in cell assay (EC<sub>50</sub> = 6 nM in Huh-7<sup>ACE2</sup> infected cells), and high selectivity against other human proteases. Crystal structure of *coronastat* in complex with the target protein allowed a better understanding of the drastic improvement in potency compared to the lead GC-376 (PDB code: 7TIZ): the trifluoromethyl group on the benzyl moiety provides two additional H-bonds with Asn<sup>142</sup>, confirming the importance of the halogen substitution for the enhancing interactions of the ligand-protein complex. These encouraging results prompted the researchers to perform further *in vivo* tests (mouse models) to evaluate both the activity and pharmacokinetic profile: *coronastat* showed no significant toxicity, high metabolic stability and high concentration in plasma and lung after both oral and intraperitoneal administration [134].

Analysing the crystal structures of M<sup>PRO</sup> in complex with *boceprevir* and *telaprevir*, two drugs with an α-ketoamide group as warhead (see section 3.2.1.2.), led to observation that the bicyclic proline can be suitably accommodated into the S2 pocket of the target protein. The merging of the pharmacophore moieties of GC-376, *boceprevir*, and *telaprevir* led to the synthesis of a new series of small molecules with inhibition effect on SARS-CoV-2 M<sup>PRO</sup> (Figure 24a). The new derivatives were characterized by: the aldehydic P1' warhead; the γ-lactam as P1 (present in GC-376); (1*R*,2*S*,5*S*)-6,6-dimethyl-3-aza-bicyclo[3.1.0]hexane-2-formamide or (1*S*,3*aR*,6*aS*)-octahydrocyclopenta[*c*]-pyrrole-1-formamide as P2 (from *boceprevir* and *telaprevir*, respectively); and a hydrophobic aromatic group as P3. MI-09, MI-23 and MI-30 (Figure 24a), tested in the M<sup>PRO</sup> biochemical inhibition assay, proved to be the most interesting compounds in the series, exhibiting IC<sub>50</sub> values in the nanomolar range (15.2 nM, 7.6 nM, and 17.2 nM, respectively). Using X-ray experiment, the most active inhibitor MI-23 was selected to further investigate the binding mode of this class of compounds with SARS-CoV-2 M<sup>PRO</sup>. Figure 24b shows the crystal structure of the complex MI-23/SARS-CoV-2 M<sup>PRO</sup>.

(PDB code: 7D3I), with the relevant interactions highlighted: the aldehydic carbon interacts with the SH of the catalytic Cys<sup>145</sup> at the S1' site, as expected; the P1 is inserted in the S1 region and forms several hydrogen bonds, as observed for the previous compounds; the conformational constraints of the bicyclic proline structure (P2) results in a restricted trans-exo conformation of the system, which determines a deep insertion into the hydrophobic S2 site of the target; the 1-ethyl-3,5-difluorobenzene moiety at P3 occupies the S3/S4 sub-region. Further biological studies were conducted on MI-09 and MI-30. The analysis of biological data showed no cytotoxicity against normal cells; remarkable antiviral activity in Vero E6 infected cells (EC<sub>50</sub> of 0.86 μM and 0.54 μM, respectively); acceptable PK properties and capability to prevent SARS-CoV-2 infection *in vivo* (transgenic mice expressing hACE2), and ability to ameliorate lung lesion and inflammation [135]. Moreover, starting from MI-09 and MI-30, Luo and co-workers recently built a computational protocol based on QSAR combined with molecular docking, MD simulations, and free binding energy MM/PBSA to design new derivatives with higher inhibition activity on SARS-CoV-2 M<sup>PRO</sup>. The virtual screening highlighted the importance of a bicycle moiety with a three-membered ring, preferably at P2, and bulky electronegative substituents (ethyl, -CN, -F and Br) on the phenyl group at P3 [136].

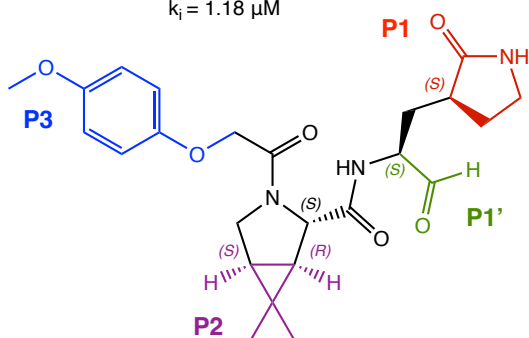
Similarly, UAWJ9-36-1 and UAWJ9-36-3, in Figure 24a, possess both the aldehyde warhead and the bicyclic fragments. *In vitro* biological studies of UAWJ9-36 derivatives showed similar inhibitory activity to GC-376 against both SARS-CoV-2 M<sup>PRO</sup> (with IC<sub>50</sub> values of 0.051 and 0.054 μM for UAWJ9-36-1 and UAWJ9-36-3, respectively), and other known human coronaviruses M<sup>PRO</sup>. Antiviral cellular assays performed on Vero E6 and Caco2-ACE2 cell lines infected with SARS-CoV-2 demonstrated the enhanced antiviral activity of UAWJ9-36-3 with EC<sub>50</sub> 0.37 and 1.06 μM against the respective cell lines. Moreover, both UAWJ9-36 derivatives showed improved selectivity against host calpains/cathepsins compared to GC-376. X-ray crystal structures of the protein co-resolved with both compounds (PDB codes: 7LYH and 7LYI, for the complex with UAWJ9-36-1 and UAWJ9-36-3, respectively), confirmed the covalent bonding and the ability of the cyclopentyl-proline and dimethyl-cyclopropyl-proline moieties to insert in the S2 site [137].

a)

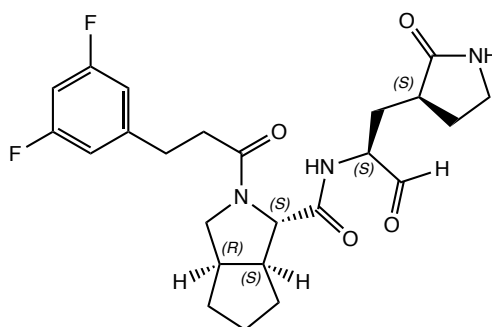


*Boceprevir*  
 $IC_{50} = 1.59 \mu M$   
 $EC_{50} = 1.90 \mu M$   
 $k_i = 1.18 \mu M$

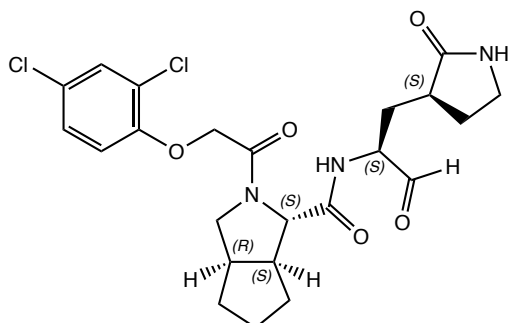
*Telaprevir*



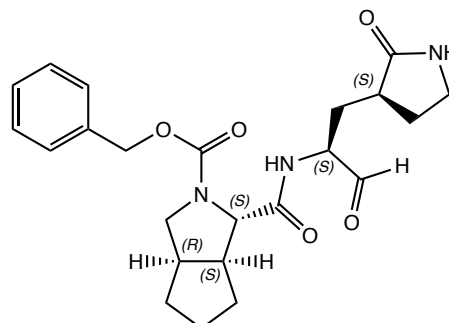
MI-09  
 $IC_{50} = 15.2 \text{ nM}$   
 $EC_{50} = 0.86 \mu M$



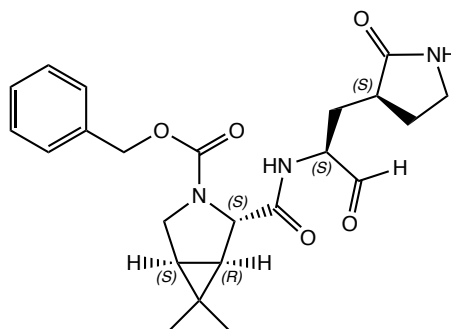
MI-23  
 $IC_{50} = 7.6 \text{ nM}$



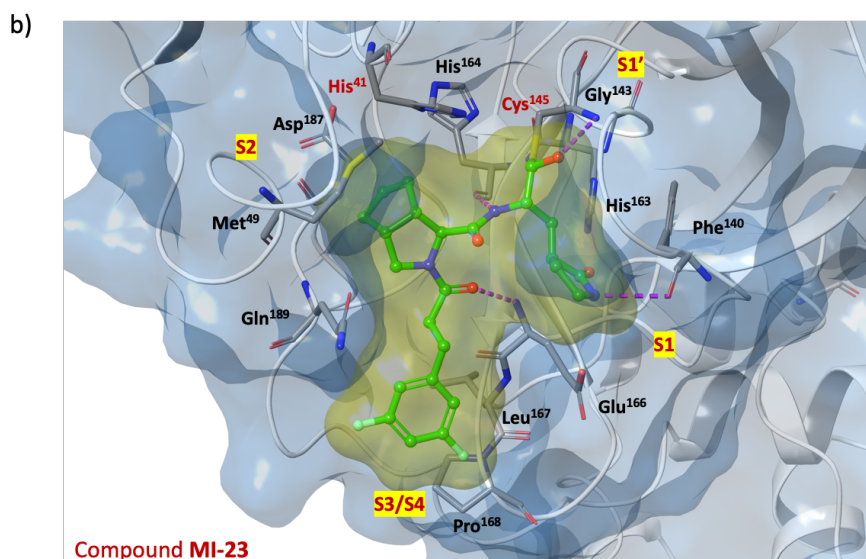
MI-30  
 $IC_{50} = 17.2 \text{ nM}$   
 $EC_{50} = 0.54 \mu M$



UAWJ9-36-1  
 $IC_{50} = 0.051 \mu M$   
 $EC_{50} = 2.56 \mu M$   
 $K_{inact}/k_i = 8.5 \times 10^4 \text{ M}^{-1}\text{s}^{-1}$



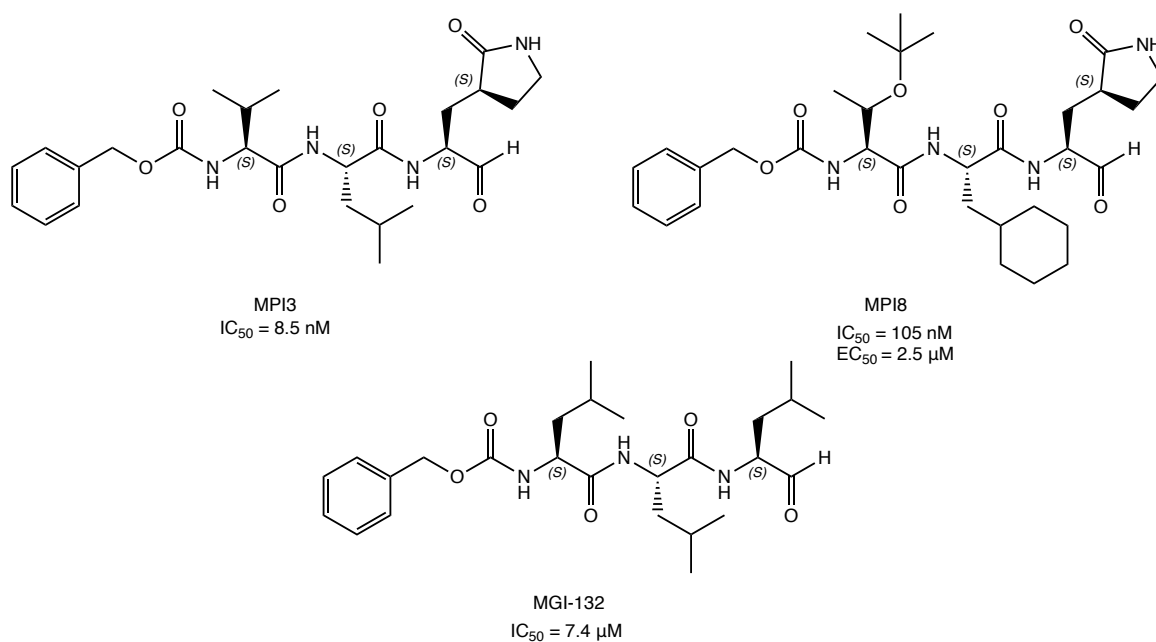
UAWJ9-36-3  
 $IC_{50} = 0.054 \mu M$   
 $EC_{50} = 0.37 \mu M$   
 $K_{inact}/k_i = 9.3 \times 10^4 \text{ M}^{-1}\text{s}^{-1}$



**Figure 24.** (a) Chemical structures of *boceprevir*, *telaprevir*, derivatives MI-09, MI-23, MI-30, and analogues UAWJ9-36-1, UAWJ9-36-3; (b) X-ray crystal structure of MI-23 in complex with SARS-CoV-2 M<sup>PRO</sup> (PDB code: 7D3I) [135,137].

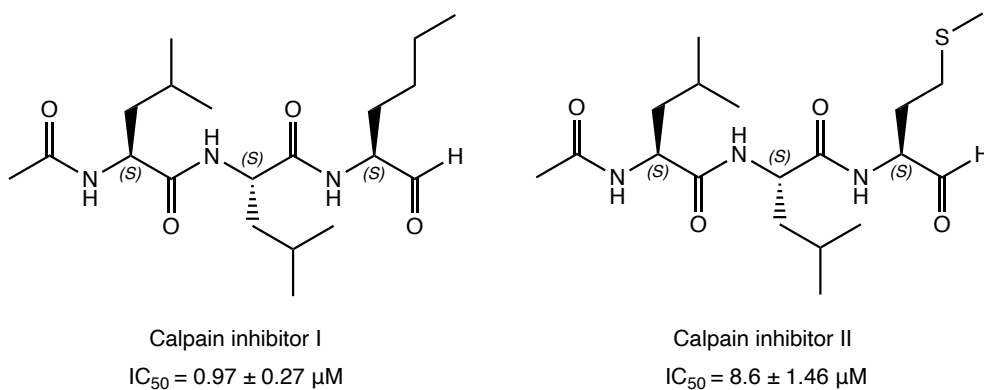
MPI3 and MPI8 (Figure 25), with the recurring 2-oxopyrrolidine side chain at the P1 site and a the Cbz group as N-terminal P3 cap, proved to be as most interesting compounds, with an IC<sub>50</sub> against SARS-CoV-2 M<sup>PRO</sup> of 8.5 nM and 105 nM, respectively. Despite the higher inhibition by MPI3, its activity in reducing viral infection in *in vitro* cell models was lower than that of MPI8, which, instead, showed completely inhibition of SARS-CoV-2 infection in Vero E6 and A549/ACE2 cell lines at concentrations of 2.5 μM and 0.31 μM, respectively. This discrepancy is probably due to the poor stability towards extra/intracellular proteases, which recognize natural amino acids (valine and leucine, in this case of MPI3) as substrates [138]. Further biological studies on MPI8 link the high potency of the tripeptidyl derivative to a dual mechanism of action: inhibition of both SARS-CoV-2 M<sup>PRO</sup> (IC<sub>50</sub> = 105 nM) and cathepsin L (IC<sub>50</sub> = 1.2 nM) [139].

Focusing on the interesting hypothesis of the dual target inhibitions of M<sup>PRO</sup> and cathepsin L, MGI-132 (Figure 25), a known proteasome inhibitor, similar to MPI3 and MPI8, showed an IC<sub>50</sub> of 7.4 μM against the M<sup>PRO</sup> and 0.15 nM against cathepsin L and a potent antiviral effect on Vero-E6 cells infected with SARS-CoV-2 [140,141].



**Figure 25.** Structures of tripeptidyl inhibitors of SARS-CoV-2 M<sup>PRO</sup>: MPI3, MPI8, and MGI-132 [138,140,141].

Similarly, through a comprehensive *in vitro* screening of investigational/approved inhibitors of the cysteinyl protease, calpain/cathepsin L covalent inhibitors I-II (Figure 26) were repurposed as potential SARS-CoV-2 M<sup>PRO</sup> inhibitors. In detail, the tripeptidyl compounds halted the enzymatic activity of SARS-CoV-2 M<sup>PRO</sup> *in vitro* at IC<sub>50</sub> values of 0.97 ± 0.27 μM and 8.6 ± 1.46 μM, respectively. Calpain inhibitor II, was able to block viral progression in the cell infection model (EC<sub>50</sub> = 2.07 ± 0.76 μM in the primary CPE assay; EC<sub>50</sub> = 3.70 ± 0.69 μM in the secondary viral yield reduction assay), with no cytotoxicity to normal cells. The analysis of the crystal structure complex of calpain inhibitor II with SARS-CoV-2 M<sup>PRO</sup> (PDB code: 6XA4) confirmed the covalent mechanism of inhibition with the formation of a hemithioacetal adduct. The methionine side chain, deeply inserted into the P1 pocket, could be the starting point for the development of new inhibitors with a dual mechanism of action (M<sup>PRO</sup> and cathepsins inhibition) [127,128,133].



**Figure 26.** Chemical structures of the repurposed calpain inhibitors I-II as new SARS-CoV-2 M<sup>PRO</sup> modulators [127,128,133].

Table 8 highlights the overall structural analysis of covalent SARS-CoV-2 M<sup>PRO</sup> inhibitors containing an aldehydic warhead, focusing on the recurrence and significance of P1, P2, and P3 portions. It is evident that the P1 portion, typically characterized by a  $\gamma$ -lactam group, plays a crucial role by forming multiple hydrogen bonds with polar residues in the S1 site, such as Phe<sup>140</sup>, His<sup>163</sup>, His<sup>164</sup>, and Glu<sup>166</sup>. This pattern is consistently observed across all the compounds listed, reinforcing the importance of maintaining this structural unit for effective interaction with the target protein.

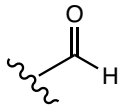
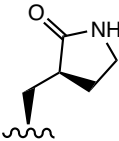
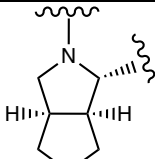
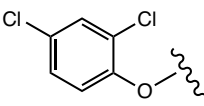
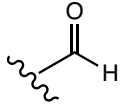
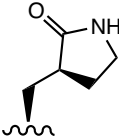
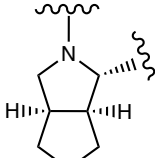
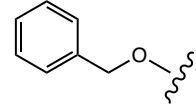
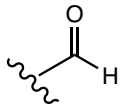
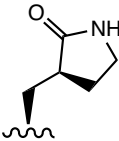
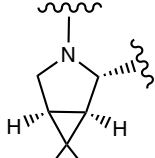
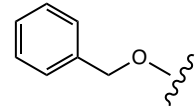
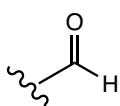
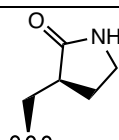
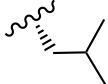

Regarding the P2 portion, a variety of hydrophobic substituents have been explored, including cycloalkyl groups (cyclohexyl for **4**, and cyclopropyl for **7** and **8**), isoleucine groups (compound GC-376, NK01-63, MPI3), aromatic rings (**5**, **6**, and UAWJ247), all capable of deeply inserting into the S2 pocket. The presence of bulky bicyclic hydrophobic groups in P2, as seen in compounds MI-09, MI-23, MI-30, UAWJ9-36-1, and UAWJ9-36-3, has been shown to enhance IC<sub>50</sub> values.

The P3 portion is particularly intriguing due to its positioning in the more exposed S3/S4 pocket, which interacts with the surface of the protease. Most of the inhibitors incorporate aromatic groups at P3, such as benzyl or phenyl substituents. Additionally, the introduction of halogen-substituted phenyl groups, as in **7**, **8**, MI-23, and MI-30, or trifluoromethyl bulky modifications, like in NK01-63, brings the P3 fragment closer to the S3/S4 cleft, enhancing hydrophobic and electrostatic interactions.

These recurring structural elements thus represent key points for the design of more potent covalent inhibitors.

**Table 8.** Analysis of structural fragments and inhibition data of the most active covalent SARS-CoV-2 M<sup>PRO</sup> inhibitors with an aldehydic warhead, by following the Schechter-Berger notation.

Cpd	P1'	P1	P2	P3	Inhibition of SARS-CoV-2 M <sup>PRO</sup> (μM)	Ref.
4					IC <sub>50</sub> = 0.053 μM	[122]
5					IC <sub>50</sub> = 0.040 μM	[122]
6					IC <sub>50</sub> = 0.034 μM	[124]
GC-376	*				IC <sub>50</sub> = 0.19 μM	[126]
7	*				IC <sub>50</sub> = 0.07 μM	[132]
8	*				IC <sub>50</sub> = 0.08 μM	[132]
UAWJ247	*				IC <sub>50</sub> = 0.045 μM	[133]
NK01-63	*				IC <sub>50</sub> = 0.016 μM	[134]
MI-09					IC <sub>50</sub> = 0.015 μM	[135]
MI-23					IC <sub>50</sub> = 0.0076 μM	[135]

<b>MI-30</b>					$IC_{50} = 0.017 \mu M$	[135]
<b>UAWJ9-36-1</b>					$IC_{50} = 0.051 \mu M$	[137]
<b>UAWJ9-36-3</b>					$IC_{50} = 0.054 \mu M$	[137]
<b>MPI3</b>					$IC_{50} = 0.0085 \mu M$	[138]

### 3.2.1.1.2. Ketone warhead

The ketone warhead has also been extensively explored in the design of SARS CoV-2 M<sup>PRO</sup> inhibitors. Among the first efforts in this direction, the  $\alpha$ -acyloxymethyl ketone **9** and the hydroxymethylketone (HMK) PF-00835231 (Figure 27), have been designed and synthesized as new M<sup>PRO</sup> SARS-CoV-2 covalent inhibitors [142]. Compound PF-00835231 exhibited an interesting  $K_i$  of 0.27 nM and an outstanding  $IC_{50}$  of 6.9 nM against the target protein. To gain insight into the binding mode and mechanism of action, PF-00835231 was co-crystallized in a covalent complex with SARS CoV-2 M<sup>PRO</sup> (Figure 27b, PDB code: 6XHM), highlighting the key contacts with the target: the carbonyl group of the HMK warhead (P1') is covalently bounded to the Cys<sup>145</sup> of the S1' site of the binding pocket, forming a tetrahedral carbinol complex, which together with the primary -OH of the HMK group reinforces the covalent interaction by additional H-bonds within the S1 cleft (His<sup>41</sup>, Gly<sup>143</sup>); the lactam P1 is positioned into the S1 pocket (as described earlier); the lipophilic leucine side chain at P2 site of the ligand is inserted into the S2 region, surrounded by hydrophobic amino acids; the indole moiety makes the complex more stable through Van der Waals interactions with the backbones of the 189-191 amino acid residues [142]. Recently, the mechanism of covalent inhibition of SARS CoV-2 M<sup>PRO</sup> was investigated by classical and QM/MM simulations, underling the importance of the hydroxymethyl group both in binding free energy increase and in the formation of the



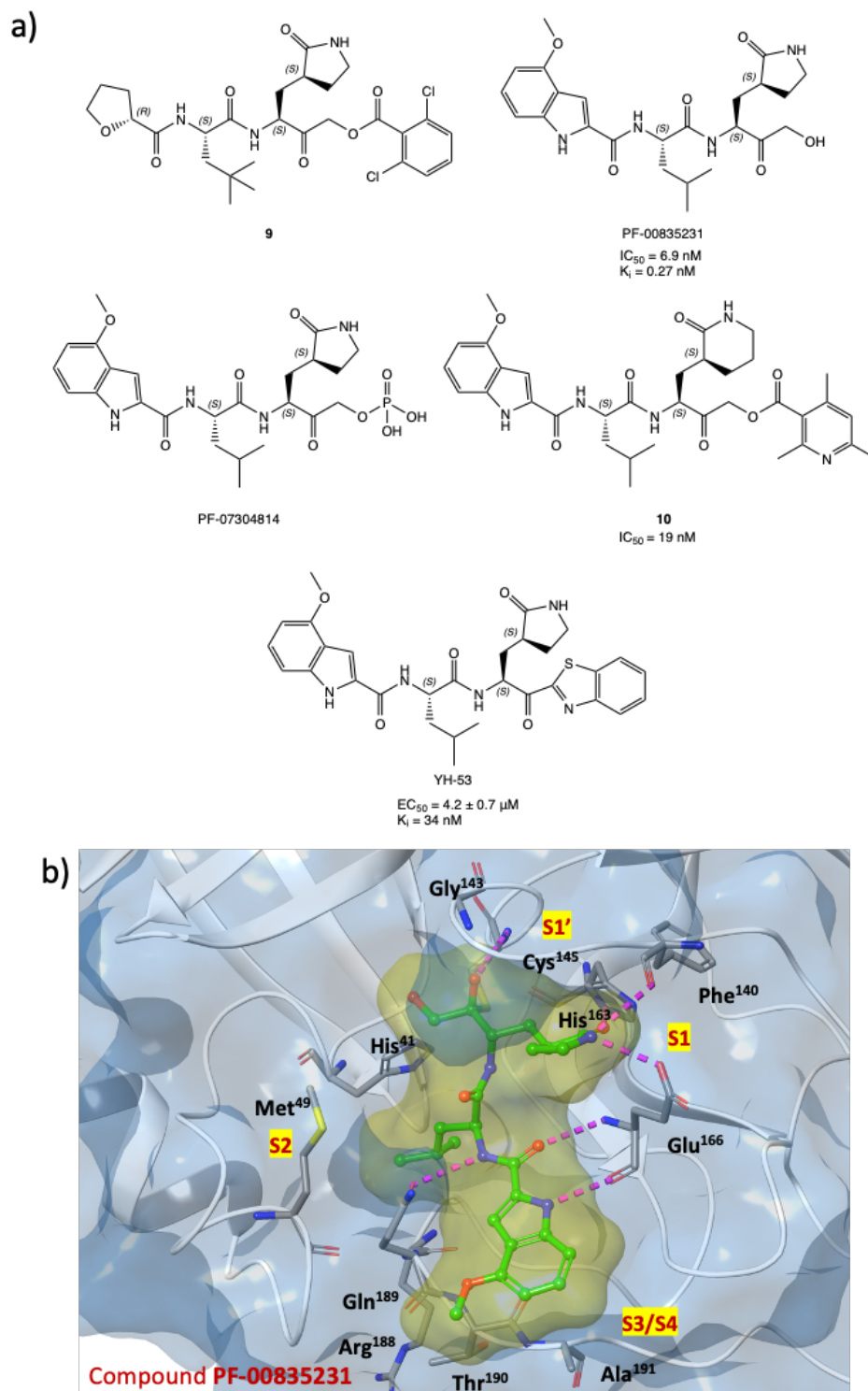
hemithioacetal adduct. Analysis of *in silico* data showed a significant inhibition effective of PF-00835231 against M<sup>PRO</sup> mutant forms.

Given the encouraging results, PF-00835231 was selected for *in vitro/vivo* biological evaluation, demonstrating: appropriate physicochemical and pharmacokinetic properties for intravenous continuous infusion; similar/higher potency compared with other peptidomimetic SARS CoV-2 M<sup>PRO</sup> inhibitors (as GC-376) or other anti-SARS CoV-2 agents (*remdesivir*); low cytotoxicity and elevated tolerability in *in vitro* model of human airway epithelium (HAEC) [142,143]. In further studies PF-00835231 evolved into the phosphate pro-drug PF-07304814 (Figure 11), which exhibited comparable antiviral and M<sup>PRO</sup> inhibitory activity but better solubility and pharmacokinetic in *in vivo* animal models; these more favourable features supported the progression of PF-07304814 in clinical trials (ClinicalTrials.gov Identifiers: NCT04627532 and NCT04535167) [144].

In order to explore the range of substitution tolerated to maintain M<sup>PRO</sup> inhibitory activity, a new series of the  $\alpha$ -acyloxymethyl ketones has been obtained by replacing the pentacyclic glutamine mimic lactam with a six-membered one and by inserting a heteroaromatic portion as an acyloxy group. The most interesting compound, **10** (Figure 27a), with a pyridyl moiety, showed a notable inhibition of M<sup>PRO</sup> (IC<sub>50</sub> = 19 nM), excellent SARS-CoV-2 replication inhibition and remarkable plasma and metabolic stability. The capability of **10** to covalently block the target has been confirmed by co-crystallization experiment (PDB code: 7MBI) [145].

With a unique P1'-benzothiazolyl ketone moiety, YH-53 (Figure 27a), previously developed as a SARS-CoV-1 M<sup>PRO</sup> blocker, also proved capable of reversibly inhibiting SARS-CoV-2 M<sup>PRO</sup>, with a K<sub>i</sub> of 34 nM and to halt SARS-CoV-2 infection *in vitro* (EC<sub>50</sub> = 4.2 ± 0.7  $\mu$ M), with high safety index and low cytotoxicity. X-ray experiments performed with the complex YH-53-SARS-CoV-2 M<sup>PRO</sup> (PDB code: 7E18) confirmed the formation of the covalent adduct and the ability of the heteroatoms of the benzothiazolyl fragment at P1' to form additional interactions within the catalytic binding site.

*In vivo* Pharmacokinetic studies in animal models highlighted that the introduction of the lipophilic benzothiazole group particularly affected the cLogP value (2.37), suggesting high cell penetration in the GI tract and excellent permeability. In contrast to the almost complete absorption, the recorded bioavailability was very low (3.6%); this issue has been ascribed to the pseudopeptide structure, responsible for high first-pass metabolism in the intestine and liver after oral administration [146,147].

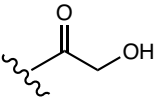
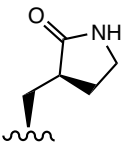
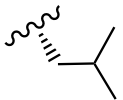
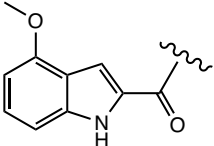
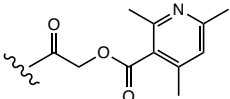
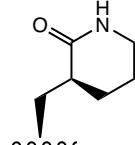
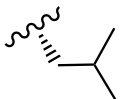
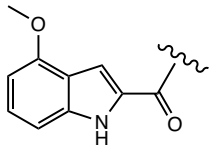
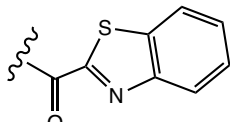
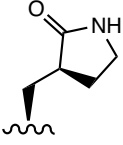
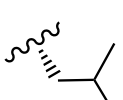
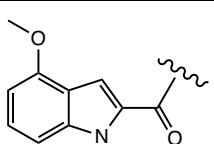


**Figure 27. (a)** Chemical structures of a series of acyloxymethyl ketone derivatives with SARS-CoV-2  $M^{\text{PRO}}$  inhibitory activity; **(b)** PF-00835231 in complex with SARS-CoV-2  $M^{\text{PRO}}$  (PDB code: 6XHM) [142,144-147].

Table 9 lists the SARS-CoV-2  $M^{\text{PRO}}$  inhibitory activity and the main structural moieties (P fragments) of the described compounds, containing a ketone as electrophilic warhead. Here, the recurrence of a  $\gamma$ -lactam group is highlighted, with the exception for compound **10**.

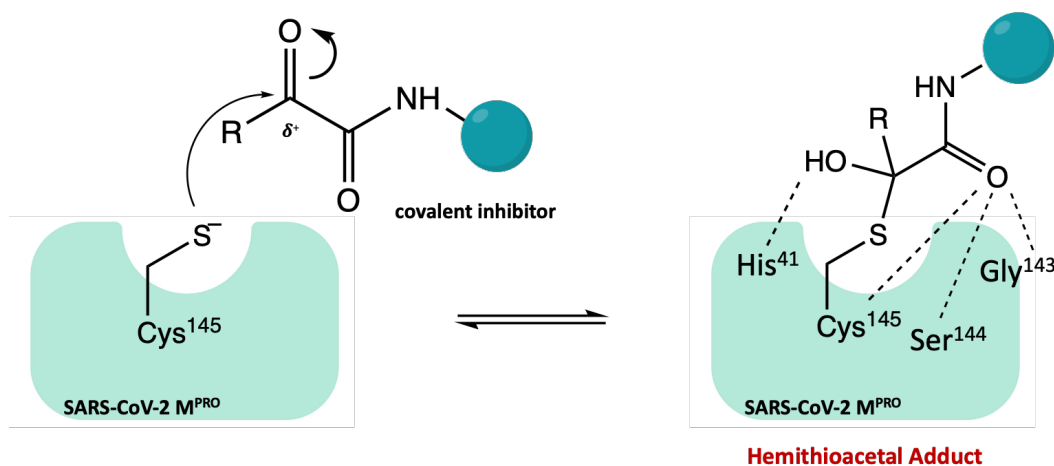
Regarding the P2 and P3 moieties, the cyclopropyl group and the indole ring, respectively, appear to be unique for the three analyzed compounds.

**Table 9.** Analysis of structural fragments and inhibition data of the most active covalent SARS-CoV-2 M<sup>PRO</sup> inhibitors with a keton warhead, by following the Schechter-Berger notation.

Cpd	P1'	P1	P2	P3	Inhibition of SARS-CoV-2 M <sup>PRO</sup> ( $\mu\text{M}$ )	Ref.
PF-00835231					IC <sub>50</sub> = 0.0069 $\mu\text{M}$	[142]
10					IC <sub>50</sub> = 0.019 $\mu\text{M}$	[145]
YH-53					K <sub>i</sub> = 0.034 $\mu\text{M}$	[146]

### 3.2.1.2. $\alpha$ -ketoamide warhead

Among the  $\alpha$ -ketoamide inhibitors, the carbonyl moiety is also crucial for the formation of a reversible covalent bond with the key cysteine residue of SARS-CoV-2 M<sup>PRO</sup>. In addition to the covalent adduct with the catalytic Cys<sup>145</sup>, the  $\alpha$ -ketoamide moiety forms additional non-covalent H-bonds with the amino acids of the active site via the carbonyl oxygen and the -OH of the hemithioacetal [148]. In detail, the nucleophilic attack by the cysteinyl -SH on the  $\alpha$ -carbonyl leads to the formation of a hemithioacetal adduct. The hydroxy group of the hemithioacetal gives a hydrogen bond to His<sup>41</sup>, while the oxygen of the carboxamide moiety accepts hydrogen bonds from the main-chain amides of Gly<sup>143</sup>, Cys<sup>145</sup> and Ser<sup>144</sup> (Figure 28) [65,149,150].



**Figure 28.** General mechanism of action of SARS-CoV-2 M<sup>PRO</sup> α-ketoamide inhibitors.

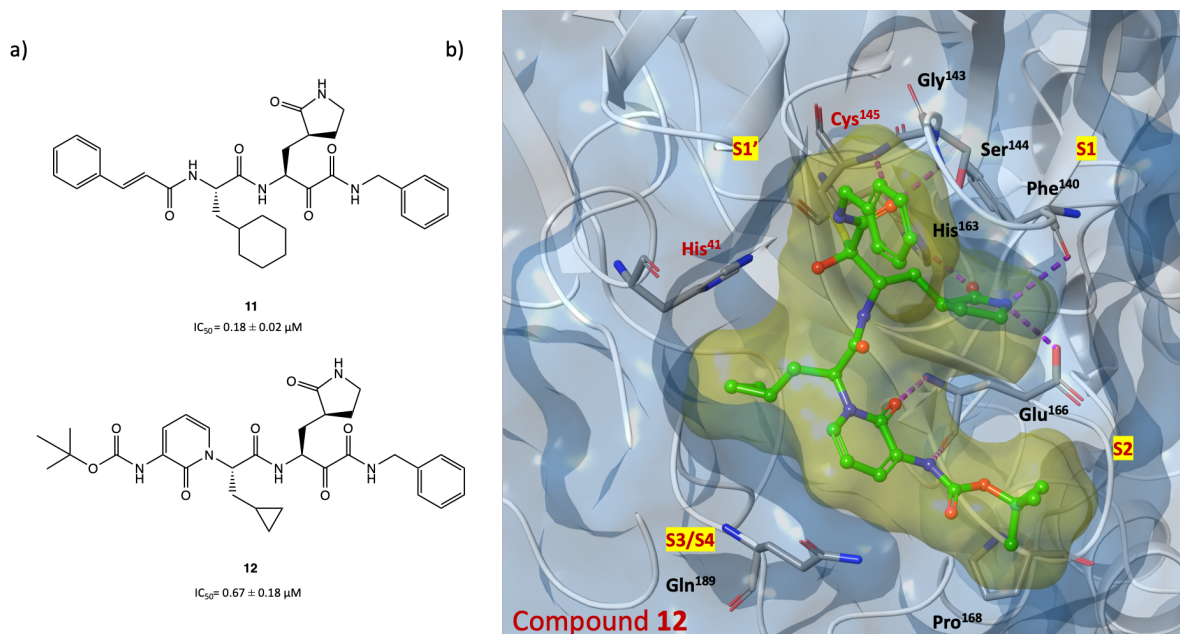
In accordance with the potential covalent inhibitory effect, the α-ketoamide warhead has been extensively used to develop new SARS-CoV-2 M<sup>PRO</sup> inhibitors.

A new promising series of compounds with a broad-spectrum antiviral activity consists of: an α-ketoamide moiety as a warhead; a 5-membered γ-lactam ring at the P1 site (also described for aldehyde and ketone warheads); a hydrophobic/alkylic group at the P2 site, such as cyclopropyl or cyclohexyl. Studies on viral protease inhibition and antiviral activity have shown that compound **11** (Figure 29a), a previously developed SARS-CoV-1 M<sup>PRO</sup> inhibitor (IC<sub>50</sub>=0.71 ± 0.36 μM), can also effectively inhibit the SARS-CoV-2 M<sup>PRO</sup> (IC<sub>50</sub> = and 0.18 ± 0.02 μM) [65,148].

Starting from the α-ketoamide **11**, as a lead compound, a lead optimization process led to the SARS-CoV-2 M<sup>PRO</sup> selective inhibitor **12** (Figure 29a). Aiming to improve the pharmacokinetic parameters, a pyridone ring has been inserted and the cinnamoyl and the cyclohexyl moieties have been replaced with a hydrophobic tert-butyloxycarbonyl and a smaller cyclopropyl groups, respectively [65]. *In vitro* biochemical assays showed that compound **12** selectively inhibited SARS-CoV-2 M<sup>PRO</sup> with IC<sub>50</sub>= 0.67 ± 0.18 μM and blocked SARS-CoV-2 viral infection in human Calu-3 cells with remarkable EC<sub>50</sub> values in the range of 4-5 μM. Moreover, *in vivo* studies conducted in CD-1 mice highlighted favorable pharmacokinetic properties and positive tropism of the compound in the lung, the primary target of COVID-19 [65].

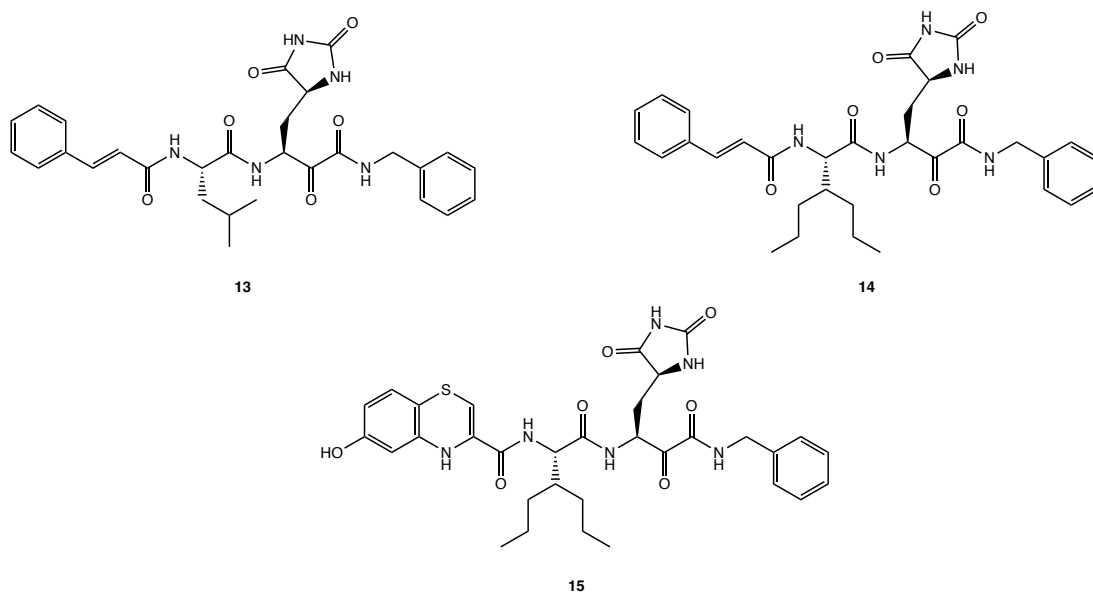
Further *in silico* studies (molecular docking and molecular dynamics simulations) showed that **12** is able to bind efficiently to the catalytic site of the protease. In Figure 29b (crystal structure of SARS-CoV-2 M<sup>PRO</sup> in complex with compound **12**, PDB code: 6Y2F), the X-ray experiment confirms the formation of the covalent adduct: the α-ketoamide group enhances the

interaction in the S1' cleft thanks to the H-bonds with His<sup>41</sup>, Gly<sup>143</sup>, Ser<sup>144</sup>, Cys<sup>145</sup>; the S1 site accommodates the glutamine surrogate  $\gamma$ -lactam ring, the key fragment amidopyridone interacts with the Glu<sup>166</sup> via two H-bonds, and the tert-butyl moiety is inserted into the S2 pocket [65,151,152].



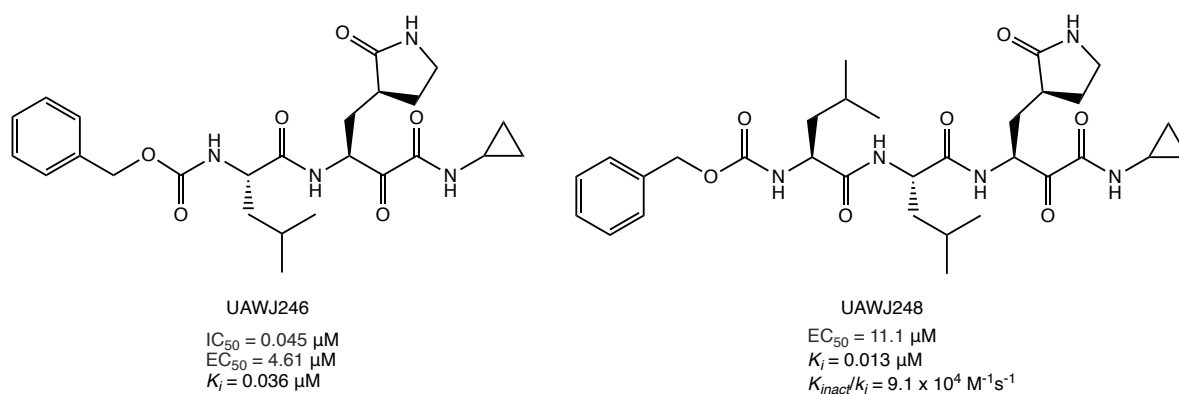
**Figure 29. (a)** Chemical structure of  $\alpha$ -ketoamide derivatives **11** and **12**; **(b)** X-ray structure of SARS-CoV-2 M<sup>PRO</sup> in complex with compound **12** (PDB code: 6Y2F) [65].

In addition, an *in silico* structural optimization study starting from derivative **11** led to compounds **13-15**, replacing the lactam moiety with an hydantoin one in P1, which is responsible for increased predicted binding affinity with SARS-CoV-2 M<sup>PRO</sup>, due to the additional H-donor/acceptor bonds (Figure 30) [153].



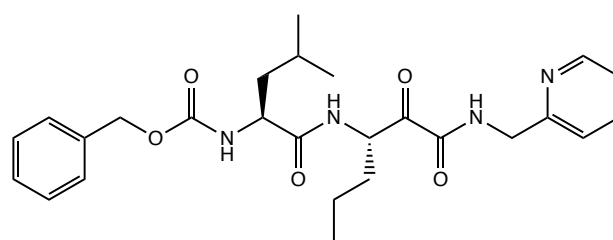
**Figure 30.** Chemical structures of hydantoin derivatives **13-15** [153].

Furthermore, starting with GC-376 as lead compound, two  $\alpha$ -ketoamide analogues: UAWJ246 and UAWJ248 (Figure 31) were investigated to examine the effects of the aldehyde warhead replacement on binding affinity with SARS-CoV-2 M<sup>PRO</sup>. Biological *in vitro* screening showed comparable inhibitory activity against M<sup>PRO</sup> with the parent compound (IC<sub>50</sub> value of 0.045  $\mu$ M for UAWJ246) [133].



**Figure 31.** Chemical structures of GC-376  $\alpha$ -ketoamide analogues UAWJ246 and UAWJ248 [133].

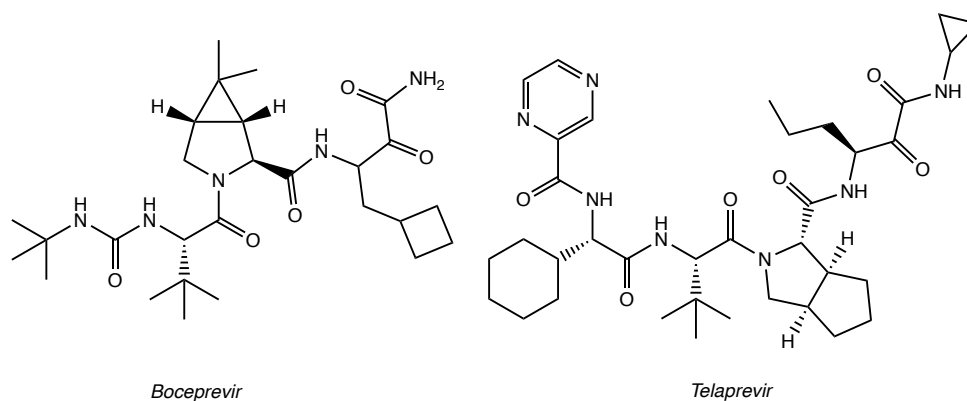
Compared to calpain inhibitors I-II, calpain inhibitor XII, in Figure 32, showed higher antiviral activity in cellular model ( $EC_{50} = 0.49 \pm 0.18 \mu M$  in the primary CPE assay and  $EC_{50} = 0.78 \pm 0.37 \mu M$  in the secondary viral yield reduction assay) and interesting SARS-CoV-2 M<sup>PRO</sup> inhibitory activity in the submicromolar range ( $IC_{50} = 0.45 \mu M$ ) [127,133,137].



calpain inhibitor XII  
 $IC_{50} = 0.45 \mu\text{M}$

**Figure 32.** Chemical structure of calpain inhibitor XII [127,133,137].

On the other hand, *boceprevir* and *telaprevir* (Figure 33), with  $\alpha$ -ketoamide warhead, proved to be interesting as antiviral compounds. Both X-ray experiments (PDB code: 6ZRU and 6ZRT) and docking studies confirmed the ability of the two drugs to form covalent adducts with Cys<sup>145</sup> [154-156]. *In vitro* biochemical assays showed significant activity of *boceprevir*, against SARS-CoV-2 M<sup>PRO</sup> with an  $IC_{50}$  of 1.59  $\mu\text{M}$  [154].



**Figure 33.** FDA-approved  $\alpha$ -ketoamide drugs: *boceprevir* and *telaprevir*.

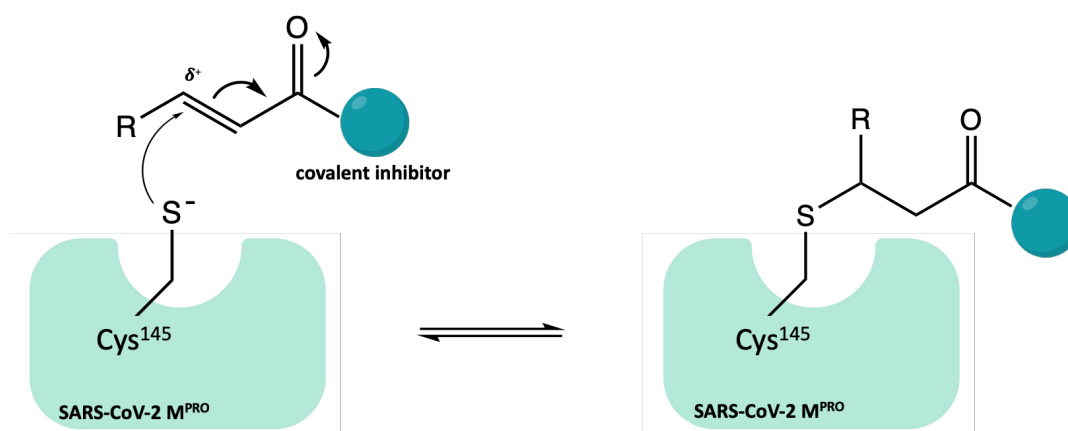
Table 10 shows the most active compounds of this class, focusing on the recurrence and significance of P1, P2, and P3 portions. As previously mentioned, the P1 moiety, which plays a key role in stabilizing the S1 site is almost always characterized by a  $\gamma$ -lactam group. For the P2 portion, a variety of hydrophobic alkyl and/or cycloalkyl substituents are common (cyclohexyl for **11**, cyclopropyl for **12**, and isoleucine for UAWJ246 and UAWJ248). Finally, most inhibitors incorporate aromatic groups at P3, except for compound **12**, which presents a *tert*-butyloxycarbonyl group. This difference may explain its higher  $IC_{50}$  value.

**Table 10.** Analysis of structural fragments and inhibition data of the most active covalent SARS-CoV-2 M<sup>PRO</sup> inhibitors with an  $\alpha$ -ketoamide warhead, by following the Schechter-Berger notation.

Cpd	P1'	P1	P2	P3	Inhibition of SARS-CoV-2 M <sup>PRO</sup> ( $\mu$ M)	Ref.
11					IC <sub>50</sub> = 0.18 $\mu$ M	[65]
12					IC <sub>50</sub> = 0.67 $\mu$ M	[65]
UAWJ246					IC <sub>50</sub> = 0.045 $\mu$ M	[133]
UAWJ248					K <sub>i</sub> = 13.20 nM	[133]

### 3.2.1.3. Michael Acceptor group as warhead

Similar to other warheads, the Michael acceptor groups inhibit the enzymes via conjugate addition of the nucleophilic cysteinyl-SH to the electrophilic C $\beta$  of the unsaturated system, producing a nearly irreversible and longer-lasting adduct. Figure 34 shows the scheme of covalent inhibition by an  $\alpha,\beta$ -unsaturated carbonyl warhead.

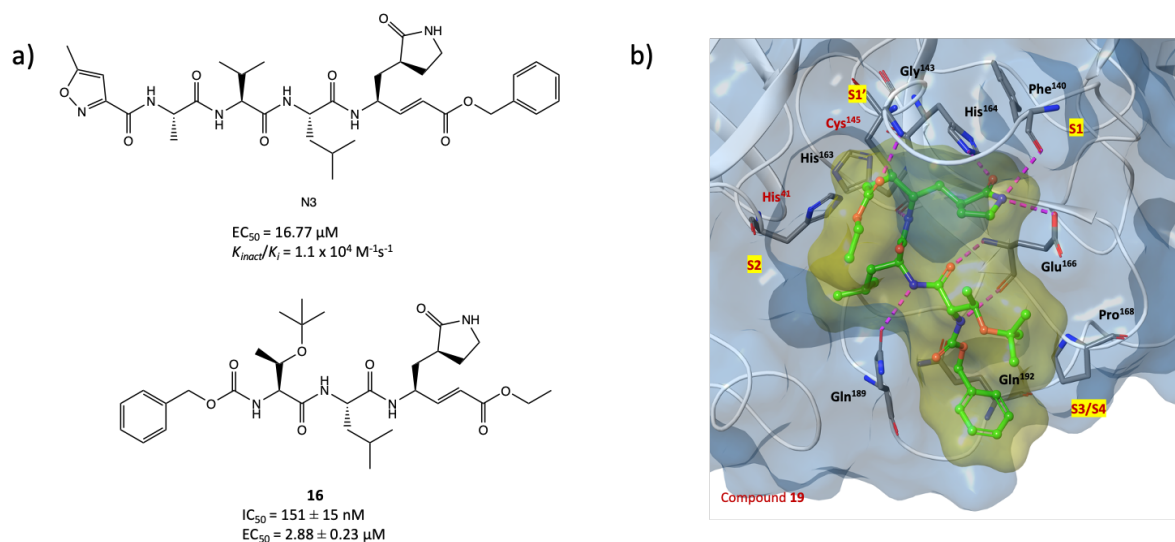


**Figure 34.** General mechanism of inhibition conjugate systems ( $\alpha,\beta$ -unsaturated carbonyl warhead is reported as an example).



N3, a compound previously evaluated for the activity against SARS-CoV-1 and MERS-CoV M<sup>PRO</sup> (Figure 35a), is one of the first peptidomimetic covalent inhibitors of SARS-CoV-2 M<sup>PRO</sup> with an  $\alpha,\beta$ -unsaturated Michael acceptor group as a warhead. Kinetics and X-ray studies explained the ability of the compound to irreversibly bind the thiol group of Cys<sup>145</sup> in the M<sup>PRO</sup> catalytic site, reacting via a conjugate addition mechanism (PDB code: 6LU7).<sup>[70]</sup> To better understand the irreversible inhibition of N3 at the atomic level, hybrid QM/MM free energy methods were also performed. In detail, the mechanism of action of N3 consists of two steps: formation of the high-energy ion pair Cys<sup>145</sup><sup>-</sup>/His<sup>41</sup><sup>+</sup> dyad and subsequent establishment of covalent bonds [157].

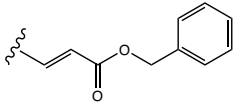
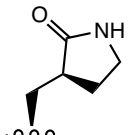
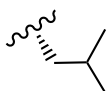
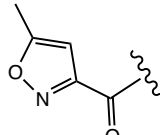
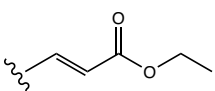
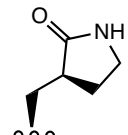
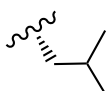
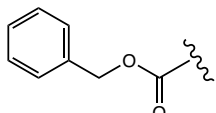
The N3 analogue **16** (Figure 35a) was identified by *in vitro* screening of previously described SARS-CoV-1 M<sup>PRO</sup> inhibitors, with promising SARS-CoV-2 M<sup>PRO</sup> inhibitory activity (IC<sub>50</sub> of 151 ± 15 nM), ability to block viral infection in Vero-E6 cells (EC<sub>50</sub> = 2.88 ± 0.23 μM), and covalent inhibition of the target protein. As illustrated by the X-ray image of the complex **16**/M<sup>PRO</sup> (Figure 35b, PDB code: 7JT7), the Sulphur atom of Cys<sup>145</sup> forms a covalent bond with the C $\beta$  atom of the vinyl group (Michael addition to the  $\alpha,\beta$ -unsaturated system), while the other portions: the  $\gamma$ -lactam ring, the isoleucine and the phenyl are inserted into protein regions S1, S2 and S4, respectively [158].



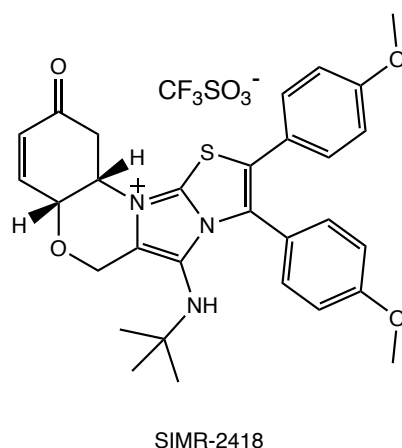
**Figure 35. (a)** Chemical structures of N3 and analogues [70,71,158,159]; **(b)** X-ray crystal structure of **16** in complex with SARS-CoV-2 M<sup>PRO</sup> [158].

Table 11 summarizes P1, P2, and P3 moieties for compounds N3 and **16** as Michael Acceptor inhibitor with a peptidomimetic scaffold.

**Table 11.** Analysis of structural fragments and inhibition data for compounds **N3** and **16**, by following the Schechter-Berger notation.

Cpd	P1'	P1	P2	P3	Inhibition of SARS-CoV-2 M <sup>PRO</sup> (μM)	Ref.
<b>N3</b>					N.A. °	[71,157]
<b>16</b>					IC <sub>50</sub> = 151 nM	[158]

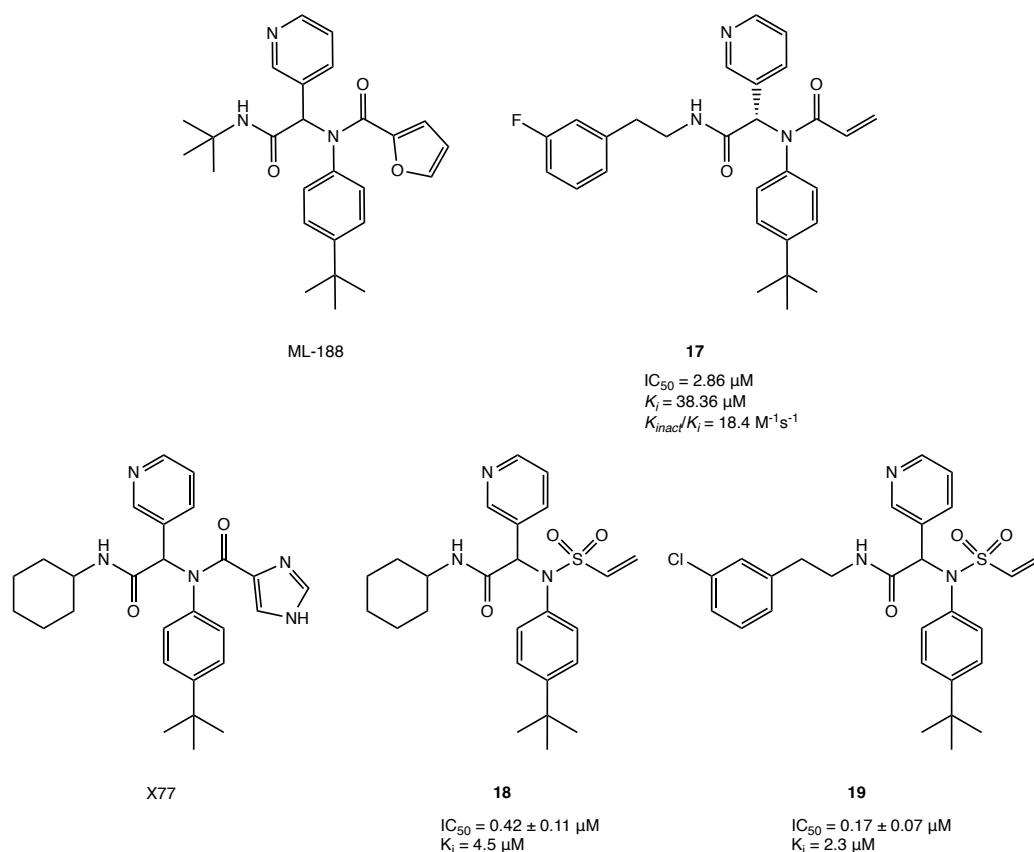
Recently, however, many non-peptido-mimetic compounds have also been proposed as SARS-CoV-2 M<sup>PRO</sup> inhibitors, leading to new ideas for exploring new chemical spaces and nuclei. An *in vitro* pilot screening of a highly focused compound library allows to identify new non-peptido-mimetic scaffolds for covalent inhibition of M<sup>PRO</sup>. The derivative SIMR-2418 proved to be one of the most active inhibitors (Figure 36), with an IC<sub>50</sub> of 20.7 μM against SARS-CoV-2 M<sup>PRO</sup> and the most favorable ADME properties. Molecular docking and dynamics analyses showed that the unexplored core, based on the fusion between a benzo[*b*][1,4]oxazin-6(5H)-one and an imidazo[2,1-*b*]thiazole system, is crucial to achieve the desired orientation in the active site of the target protein. The presence of a cyclohexenedione fragment is essential for the covalent inhibition, while the *t*-butyl group fits deeply into a hydrophobic cleft[160].



**Figure 36.** SIMR-2418, a promising non-peptidomimetic SARS-CoV-2 M<sup>PRO</sup> inhibitor [160].

Interestingly, a new covalent SARS CoV-2 M<sup>PRO</sup> inhibitor with a Michael acceptor group has been discovered by using a recently developed automated pipeline, called *covalentizer*. This innovative computational protocol allowed to design a series of irreversible acrylamide agents, starting from the reversible non-peptido-mimetic SARS CoV-1/2 M<sup>PRO</sup> inhibitor ML188. The most interesting compound, **17** (S-enantiomer, Figure 37), showed a significant IC<sub>50</sub> of 2.86 μM (the R-enantiomer was nearly inactive, with an IC<sub>50</sub> of 86.32 μM). The X-ray crystal structure (PDB code: 7NW2) provided important details about the binding mode of **17** in the catalytic pocket of SARS CoV-2 M<sup>PRO</sup>: the conjugate acrylamide moiety, which replaces the furan one, is reactive to the nucleophilic conjugate addition of the SH; the p-tert-butylphenyl moiety is projected deeply into the hydrophobic S2 pocket; the 3-F-phenethylamide fragment, absent in the lead ML-118, is essential to establish key contacts with the S4 cleft [161].

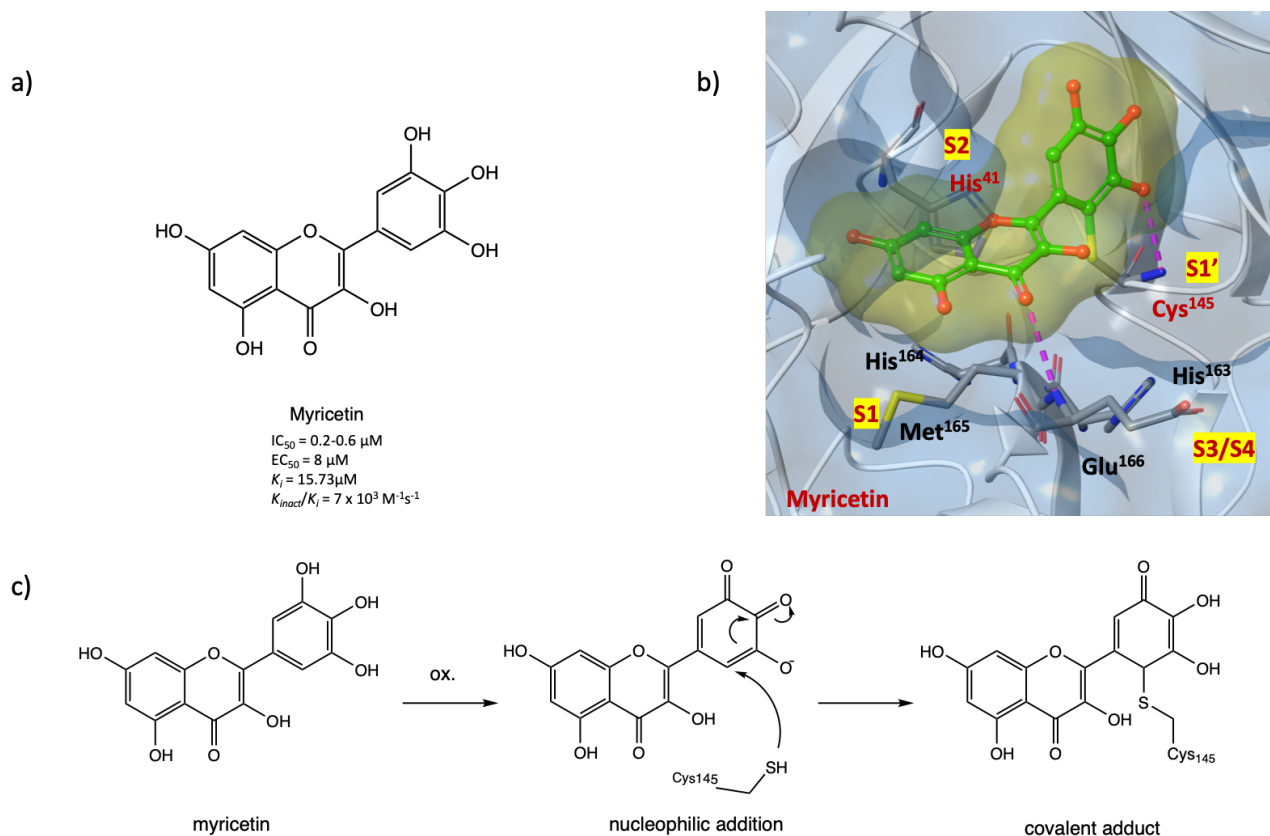
Similarly, starting from the analysis of a previously developed non-covalent SARS CoV-2 M<sup>PRO</sup> inhibitor X77, in Figure 37, the replacement of its imidazole moiety, located near the catalytic Cys<sup>145</sup> (PDB code: 6W63), with several conjugated warheads has been investigated. One of the most interesting compounds, **18** (Figure 37), with an unusual vinyl sulfone moiety, showed an IC<sub>50</sub> of 0.42 ± 0.11 μM against SARS CoV-2 M<sup>PRO</sup>, an order of magnitude stronger than X77 (IC<sub>50</sub> = 4.1 μM). The covalent inhibition mechanism was confirmed by both kinetic (ITC, isothermal titration calorimetry) and crystallographic analyses (PDB code: 7MLG). Further SAR studies performed with the designed lead compounds allowed the optimization of the interaction with the target; in particular, it was found that the substitution of the cyclohexyl portion with longer chains led to compounds with higher potency compared to **18** (compound **19** showed an IC<sub>50</sub> of 0.17 ± 0.07 μM, Figure 37)[162].



**Figure 37.** Examples of non-peptidomimetic SARS-CoV-2 M<sup>PRO</sup> inhibitors with a covalent mechanism of action [161,162].

In addition to the above examples, several non-peptidomimetic natural products, such as the conjugated flavonoid systems, have also emerged as interesting lead compounds for the development of novel covalent SARS CoV-2 M<sup>PRO</sup> inhibitors in extensive *in silico/in vitro* screening campaigns. *Myricetin*, with its novel hidden electrophilic pyrogallol portion (Figure 38a) is one of the best characterized [163]. The enzymatic inhibition assay confirmed its great potency with an  $IC_{50}$  of 0.2-0.6  $\mu\text{M}$  against M<sup>PRO</sup>. The solved X-ray crystal structure in complex with SARS-CoV-2 M<sup>PRO</sup> showed that the C6' of the *myricetin* is covalently linked to the catalytic Cys<sup>145</sup> (Figure 38b, PDB code: 7B3E) [141,164].

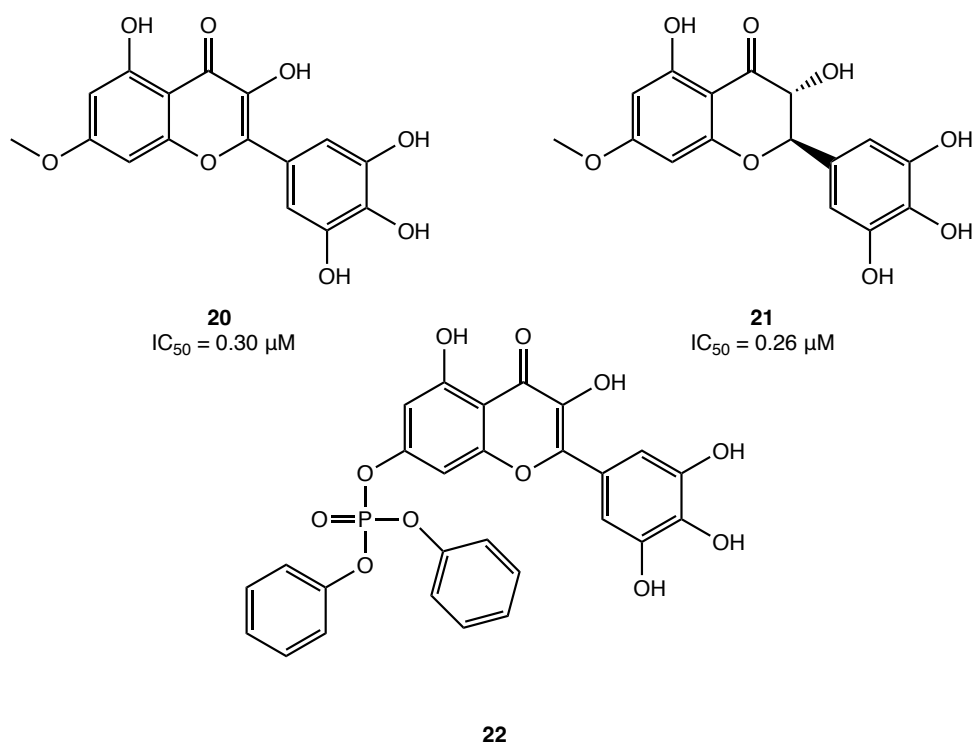
Considering the structure of *myricetin*, the irreversible inhibition could be unexplained. However, a preliminary *in vivo* oxidative step was proposed to explain the biological activity (Figure 38c): after the oxidation, the pyrogallol fragment is converted to an *o*-quinone, which could function as an  $\alpha,\beta$ -unsaturated carbonyl group (as in the *p*-quinone group of vitamin K3, see above). The sulfur atom of Cys<sup>145</sup> can attack *o*-quinone and the resulting prototropic equilibrium could lead to the formation of the covalent adduct [164].



**Figure 38.** (a) Chemical structure of the flavonoid compound *myricetin*; (b) X-ray crystal structure of the covalent complex between *myricetin* and SARS-CoV-2 M<sup>PRO</sup> (PDB code: 7B3E); (c) hypothetical mechanism of action of *myricetin* in inhibiting SARS-CoV-2 M<sup>PRO</sup> [141,164].

Further *in vitro* cellular assays highlighted the capability of *myricetin* to block SARS-CoV-2 infection in Vero E6 cells with an  $EC_{50}$  value of  $8 \mu M$  [164]. Moreover, in an *in vivo* model of lung injury in mice it was shown to inhibit the infiltration of inflammatory cells and the production of pro-inflammatory cytokines (e.g. IL-6, TNF-  $\alpha$ ), thereby alleviating the overall inflammation [165]. These findings prompted the scientific community to develop new inhibitors based on the *myricetin* scaffold.

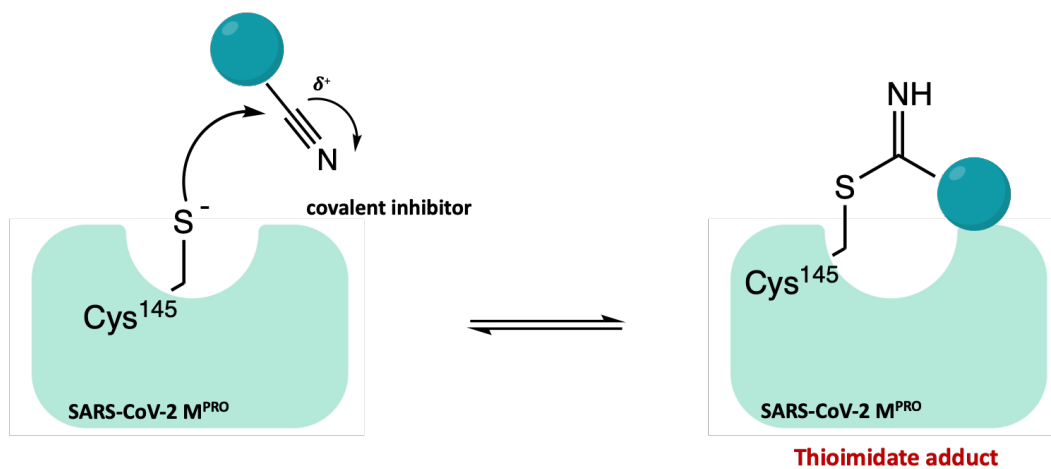
Indeed, by inserting a p-OCH<sub>3</sub> at the C7 of *myricetin* core, derivatives **20** and **21** (Figure 39) were obtained with improved potency compared with the parent compound ( $IC_{50} = 0.30$  and  $0.26 \mu M$ , respectively). In addition, the insertion of a phosphonate group to the 7-OH of *myricetin* led to compound **22** (Figure 39), which exhibited the highest inhibitory activity against SARS-CoV-2 M<sup>PRO</sup>, demonstrating the consistency of the prodrug strategy for further development [164].



**Figure 39.** Chemical structures of promising *myricetin* analogues as covalent SARS-CoV-2 M<sup>PRO</sup> inhibitors [164].

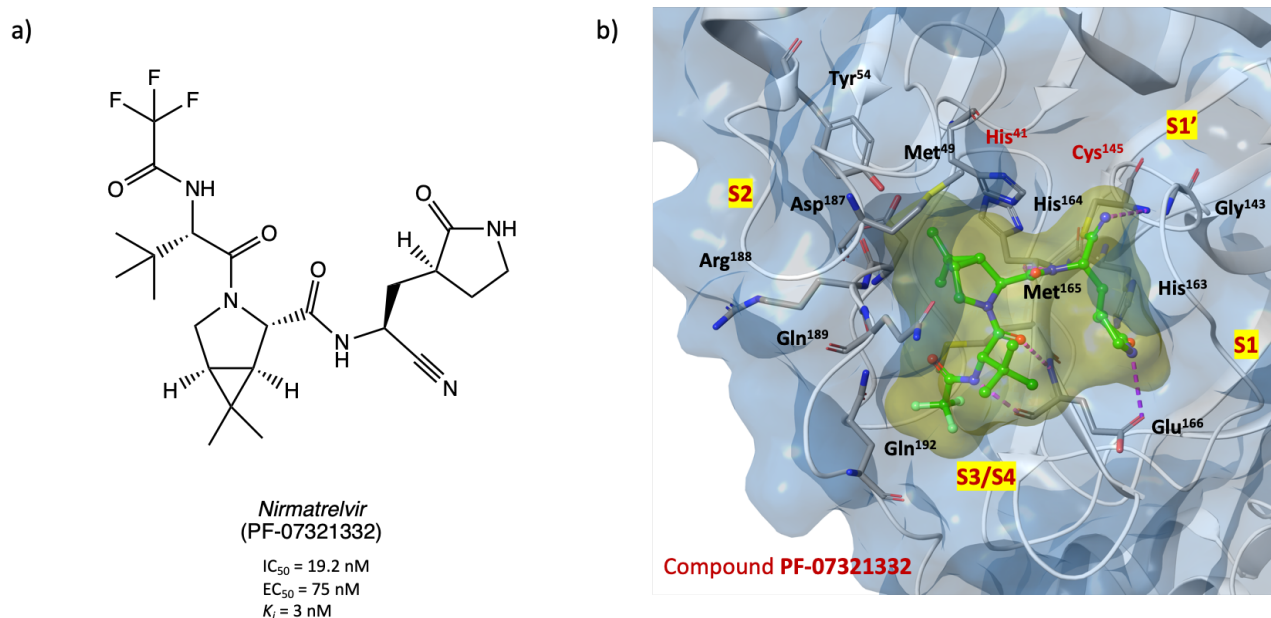
#### 3.2.1.4. Nitrile warhead

Interestingly, the nitrile group has also been proposed as remarkable warhead for the development of SARS-CoV-2 M<sup>PRO</sup> covalent inhibitors. Due to the difference in electronegativity with the nitrogen atom, the carbon of nitrile is susceptible to nucleophilic addition by the catalytic cysteine of the protease, with subsequent formation of the reversible thioimide covalent adduct, as shown in Figure 40.



**Figure 40.** General mechanism of SARS-CoV-2 M<sup>PRO</sup> inhibition with the nitrile warhead.

In this light, Pfizer has undertaken a comprehensive drug discovery campaign to develop new SARS-CoV-2 M<sup>PRO</sup> inhibitors for oral administration. Starting with the previously developed carbonyl compound PF-00835231, which exhibited potent M<sup>PRO</sup> inhibition and anti-SARS-CoV-2 activity but no intestinal absorption, Pfizer researchers developed the first orally bioavailable anti-SARS-CoV-2 compound PF-07321332 (*nirmatrelvir*) (Figure 41a), with a nitrile group as reactive warhead. In preliminary biological assays, *nirmatrelvir* showed potent SARS-CoV-2 M<sup>PRO</sup> inhibitory activity ( $K_i = 3$  nM and  $IC_{50} = 19.2$  nM) and anti-SARS-CoV-2 activity in cell-based assays ( $EC_{50} = 75$  nM). Its less peptidomimetic structure (fewer H-bond donors and lower polarity compared to the parent compound) and the insertion of a trifluoroacetamide moiety guaranteed excellent intestinal barrier permeation and excellent oral bioavailability since the first *in vivo* evaluations [166,167].



**Figure 41.** (a) Chemical structure of the orally bioavailable compound PF-07321332; (b) X-ray crystal structure of PF-07321332 in complex with SARS-CoV-2 M<sup>PRO</sup> (PDB code: 7VH8) [116].

Thanks to its remarkable results obtained in the preclinical evaluation, it is currently under phase 3 of clinical trials in combination with *ritonavir* (PAXLOVID<sup>®</sup>, see ClinicalTrials.gov Identifier: NCT04960202), an already approved anti-HIV agent insert in the formulation as pharmacokinetic enhancer (it increases the systemic exposure and the half-life of *nirmatrelvir* thanks to its inhibitory activity on the cytochrome metabolizing enzymes). The good results already obtained prompted the European Medicines Agency (EMA) and the corresponding

authorities of United States and United Kingdom to grant a conditional marketing authorization for the treatment of COVID-19 [168-170].

From a medicinal chemistry point of view, the crystal structure of SARS-CoV-2 M<sup>PRO</sup> in complex with PF-07321332 (PDB code: 7VH8 and 7VLQ) has been resolved to clarify the mechanism of action of this compound. As shown in Figure 41b, the nitrile warhead can form a reversible covalent thioimidate adduct with the sulfur atom of the Cys<sup>145</sup> at S1'; furthermore, the formed imine nitrogen enhances the reversible adduct by interacting with the Gly<sup>143</sup> and Cys<sup>145</sup> residues of the oxyanion hole. The usual  $\gamma$ -lactamic ring fits into the S1 pocket and interacts with His<sup>163</sup> and Glu<sup>166</sup>, while the dimethyl-bicycloproline group is collocated into the hydrophobic S2 pocket and is surrounded by the side chains of His<sup>41</sup>, Met<sup>49</sup>, Tyr<sup>54</sup>, Met<sup>165</sup> and Gln<sup>189</sup>, resulting in extensive Van der Waals interactions. The trifluoromethyl group, instead, is important to anchor the inhibitor at the S4 sub-pocket, by forming stabilizing contact with Gln<sup>192</sup> and two ordered small molecules positioned in this site [116,171].

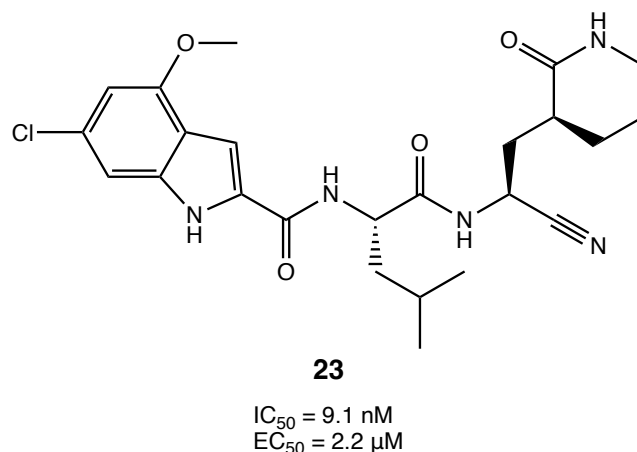
Furthermore, giving the importance of PF-07321332 as the first SARS-CoV-2 M<sup>PRO</sup> inhibitor approved for clinical use, a lot of advanced computational studies, such as steered molecular dynamics and classical-hybrid QM/MM simulations, have been conducted with the aim to clarify the mechanism of binding/inhibition and, thus, to guide the design of new analogues. It has been demonstrated as the P1 ( $\gamma$ -lactamic ring) and the P2 (dimethyl-cyclopropylproline group) results essential in the ligand-binding process, contributing positively to the total binding free energy (substitution with other groups determined a drastic reduction of binding affinity). On the other hand, it has been evidenced as the P3 and P4 groups (isobutyl and trifluoromethyl) made favorable but small contribution to the binding free energy, suggesting the possibility of modification on these sites to increase the binding strength [172,173].

In addition, given the widespread distribution of SARS-CoV-2 variants of concern (beta, delta, and the rapidly spreading omicron variant), several studies are currently being conducted to investigate the efficacy of PF-07321332 providing very encouraging results. Indeed, *nirmatrelvir* showed capability to inhibit the most prevalent M<sup>PRO</sup> variants expressed by the most diffused SARS-CoV-2 mutant lineages *in vitro* [174], and capability to potently block the infection of beta, delta, and omicron SARS-CoV-2 variants both *in vitro* and *in vivo* animal models [175-179].

Starting from the carbonyl derivatives GC-376 or PF-07304814, analogues with nitrile warhead were designed to explore the impact of this one on the inhibition activity. Among



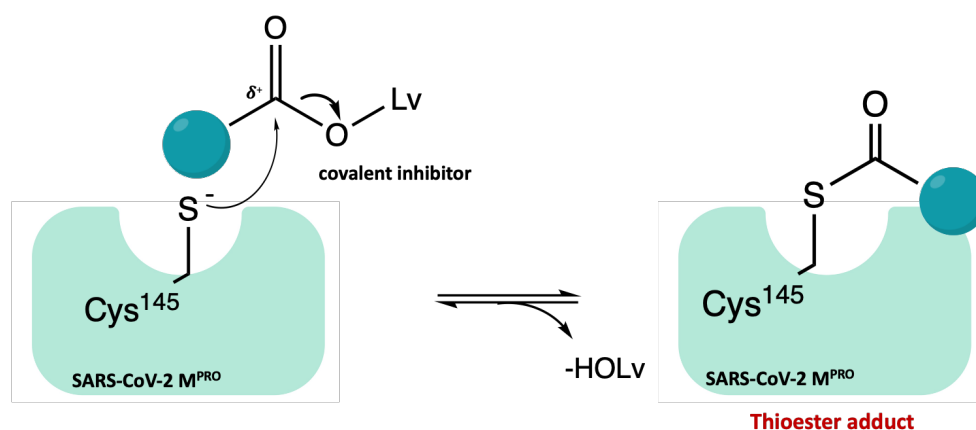
all, compound **23** (Figure 42) was the most active in the inhibition assay, with an  $IC_{50}$  of 9.1 nM against the target protein and an  $EC_{50}$  of 2.2  $\mu$ M in plaque reduction assay conducted in Vero E6 infected cells. In general, by analyzing the results for the whole series, it appeared as the substitution of the carbonyl warhead with a nitrile one affect positively the activity, leading to more active and selective compounds [180].



**Figure 42.** Chemical structure of the nitrile derivative **23** [180].

### 3.2.1.5. Ester warhead

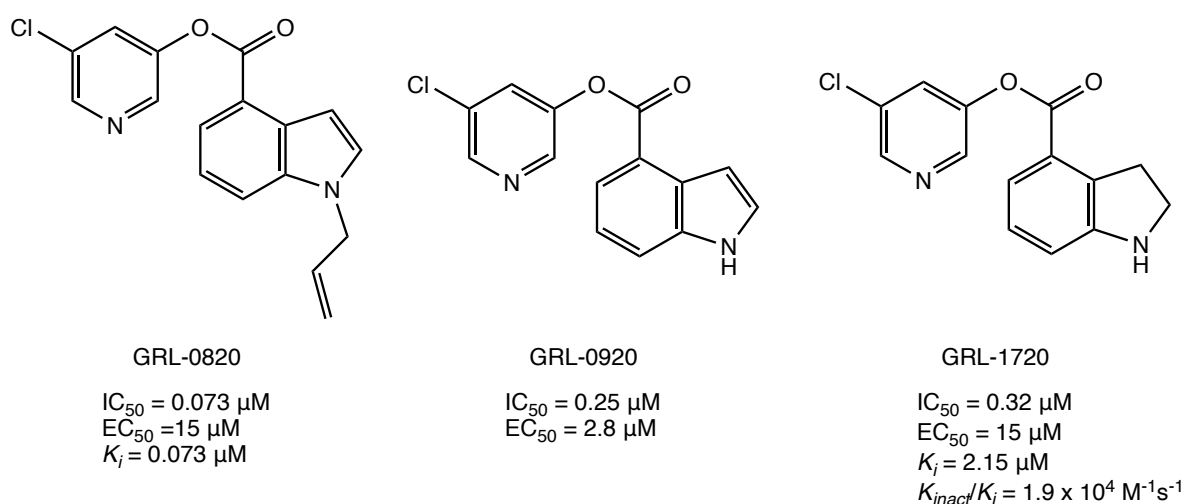
Figure 43 shows the general mechanism of action of the class of inhibitor molecules with an ester warhead, which involves a nucleophilic attack of the Cys<sup>145</sup> on the electrophilic carbonyl of the ester group, followed by cleavage of the alkoxy group (-OLv) and irreversible acylation of the enzyme.



**Figure 43.** Mechanism of acylation of SARS-CoV-2 M<sup>PRO</sup> mediated by ester derivatives.

The class of indole/indoline chloropyridinyl esters is one of the most frequently described in the literature as SARS CoV-2 M<sup>PRO</sup> inhibitors [147,181-183]. Three indole/indoline-chloropyridinyl-ester derivatives (GRL-0820, GRL-0920, and GRL-1720 in Figure 44), which

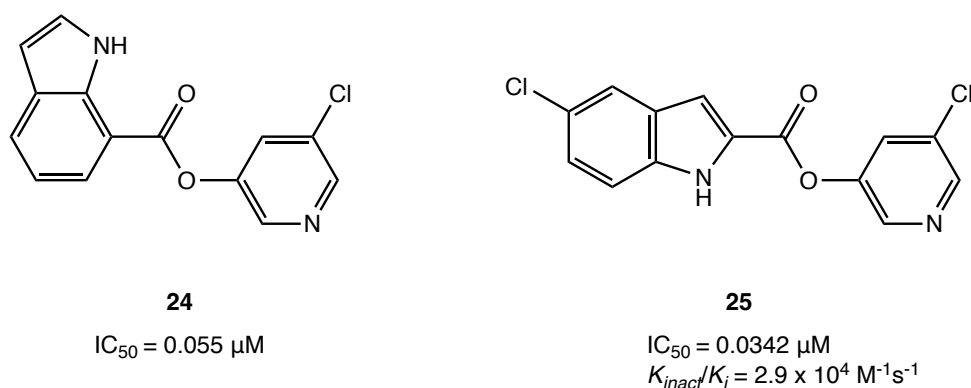
have already been evaluated against SARS-CoV-1, showed as promising lead compounds for the design of new SARS-CoV-2 M<sup>PRO</sup> inhibitors. All compounds showed potent inhibitory activity against SARS-CoV-2 M<sup>PRO</sup> (IC<sub>50</sub> values of 0.073 μM, 0.25 μM, and 0.32 μM, respectively) and anti-SARS-CoV-2 activity in Vero E6 infected cells, with EC<sub>50</sub> values of 15 μM, 2.8 μM and 15 μM, respectively [147,181,182].



**Figure 44.** Chemical structures of chloropyridinyl ester GRL-0820, GRL-0920, and GRL-1720 repurposed as SARS-CoV-2 M<sup>PRO</sup> inhibitors [147,181,182].

*In silico* docking analyses and kinetic studies proved that all three derivatives are able to covalently bind the catalytic Cys<sup>145</sup> through an acyl substitution (as described in the general mechanism of action in Figure 43) [181,182].

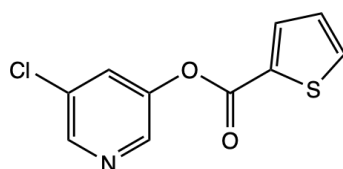
Considering that the chloro-substituted pyridinyl group is a common pharmacophoric moiety, a new series of indole esters has been investigated (Figure 45). In details, compounds **24** and **25** proved to be the most potent inhibitors with IC<sub>50</sub> values of 0.055 μM and 0.0342 μM, respectively. Further biological assays showed that both compounds could inhibit M<sup>PRO</sup> in HEK and A549 human lung epithelial cell lysate [183].



**Figure 45.** Chemical structures of indole derivatives **24** and **25** [183].

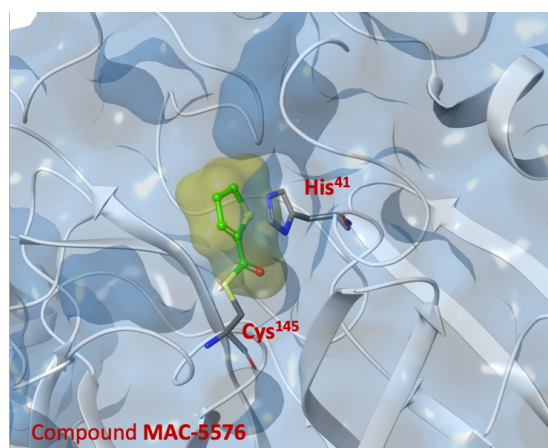
Interestingly, through a comprehensive *in vitro* screening of previously developed SARS-CoV-1 M<sup>PRO</sup> inhibitors, the chloropyridinyl ester MAC-5576 (Figure 46a), which exhibited an interesting IC<sub>50</sub> of 81 nM against SARS-CoV-2 M<sup>PRO</sup>, was identified. However, in the cytopathic reduction assay performed in Vero-E6 cells, it did not show the desired inhibition of viral infection. Nevertheless, the X-ray crystal structure of MAC-5576 in complex with the target protein has been resolved, to highlight the importance of this new non-peptidomimetic scaffold for the design of new covalent inhibitors: as shown in Figure 46b (PDB code: 7JT0), the compound is able to acylate the protease [158].

a)



MAC-5576  
IC<sub>50</sub> = 81 nM

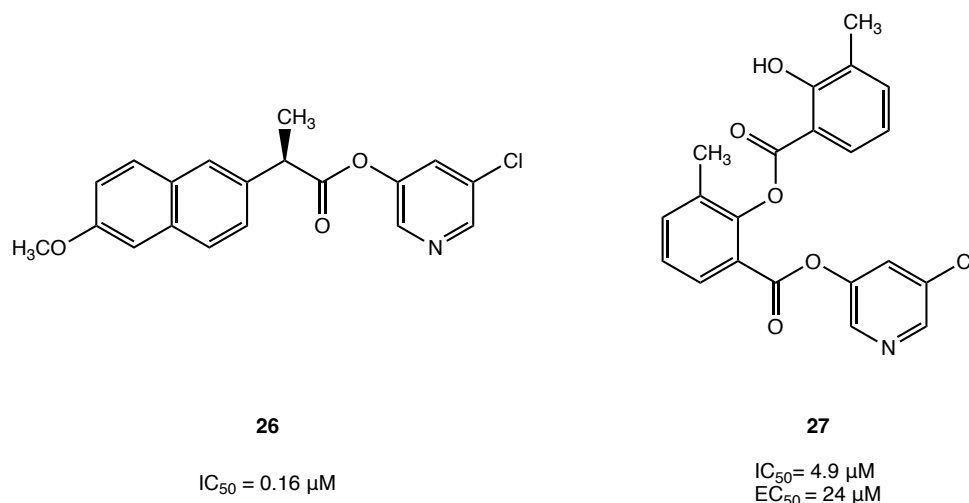
b)



**Figure 46. (a)** Chemical structure of MAC-5576, a chloropyridinyl ester SARS-CoV-2 M<sup>PRO</sup> inhibitor; **(b)** Crystal structure of the covalent complex SARS-CoV-2 M<sup>PRO</sup> with MAC-5576 (PDB code: 7JT0) [158].

The design of a series of new 5-chloropyridinyl esters of non-steroidal anti-inflammatory drugs (NSAIDs, as *salicylic acid*, *ibuprofen*, *naproxen*, *indomethacin*) led to interesting compounds. Among all of them, (*R*)-*naproxen* derivative **26** (Figure 47) with an IC<sub>50</sub> value of 0.16 μM was the most interesting compound. In addition, compound **27** (Figure 47), which

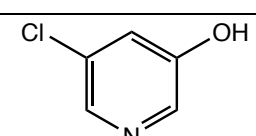
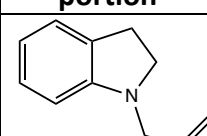
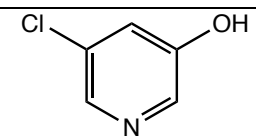
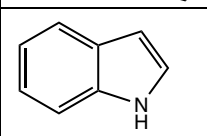
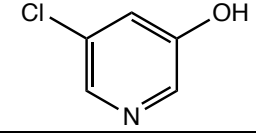
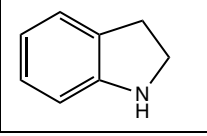
proved less efficacy in the inhibition assay ( $IC_{50} = 4.9 \mu\text{M}$ ), exhibited a potent antiviral activity in cellular assays ( $EC_{50} = 24 \mu\text{M}$  in VeroE6 cells). MALDI-TOF analysis was used to demonstrate the ability of the two compounds to covalently bind the SARS-CoV-2  $M^{\text{PRO}}$ , leading to the irreversible acylation of the enzyme [184].

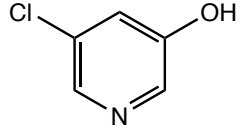
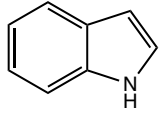
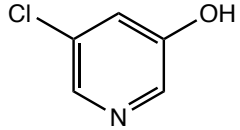
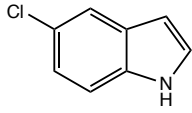


**Figure 47.** Chemical structures of 5-chloropyridinyl ester derivatives of nonsteroidal anti-inflammatory agents **26** and **27** [184].

In Summary, the presence of the chloro-substituted pyridinyl group is observed in all the analyzed compounds and represents an essential pharmacophoric moiety; furthermore, maintaining the chloro-pyridinyl portion, the class of indole esters has been extensively examined in literature as SARS CoV-2  $M^{\text{PRO}}$  inhibitors, Table 12.

**Table 12.** Analysis of structural fragments and inhibition data of the most active ester SARS-CoV-2  $M^{\text{PRO}}$  inhibitors.

Cpd	Alcoholic portion	Carboxylic portion	Inhibition of SARS-CoV-2 $M^{\text{PRO}}$ ( $\mu\text{M}$ )	Ref.
<b>GRL-0820</b>			$IC_{50} = 0.073 \mu\text{M}$	[147,181,182]
<b>GRL-0920</b>			$IC_{50} = 0.25 \mu\text{M}$	[147,181,182]
<b>GRL-1720</b>			$IC_{50} = 0.32 \mu\text{M}$	[147,181,182]

<b>24</b>			IC <sub>50</sub> = 0.055 μM	[183]
<b>25</b>			IC <sub>50</sub> = 0.0342 μM	[183]

### 3.2.1.6. Selenium/sulfur as electrophilic warhead: the case of *ebsele*/*esulfur* and analogues

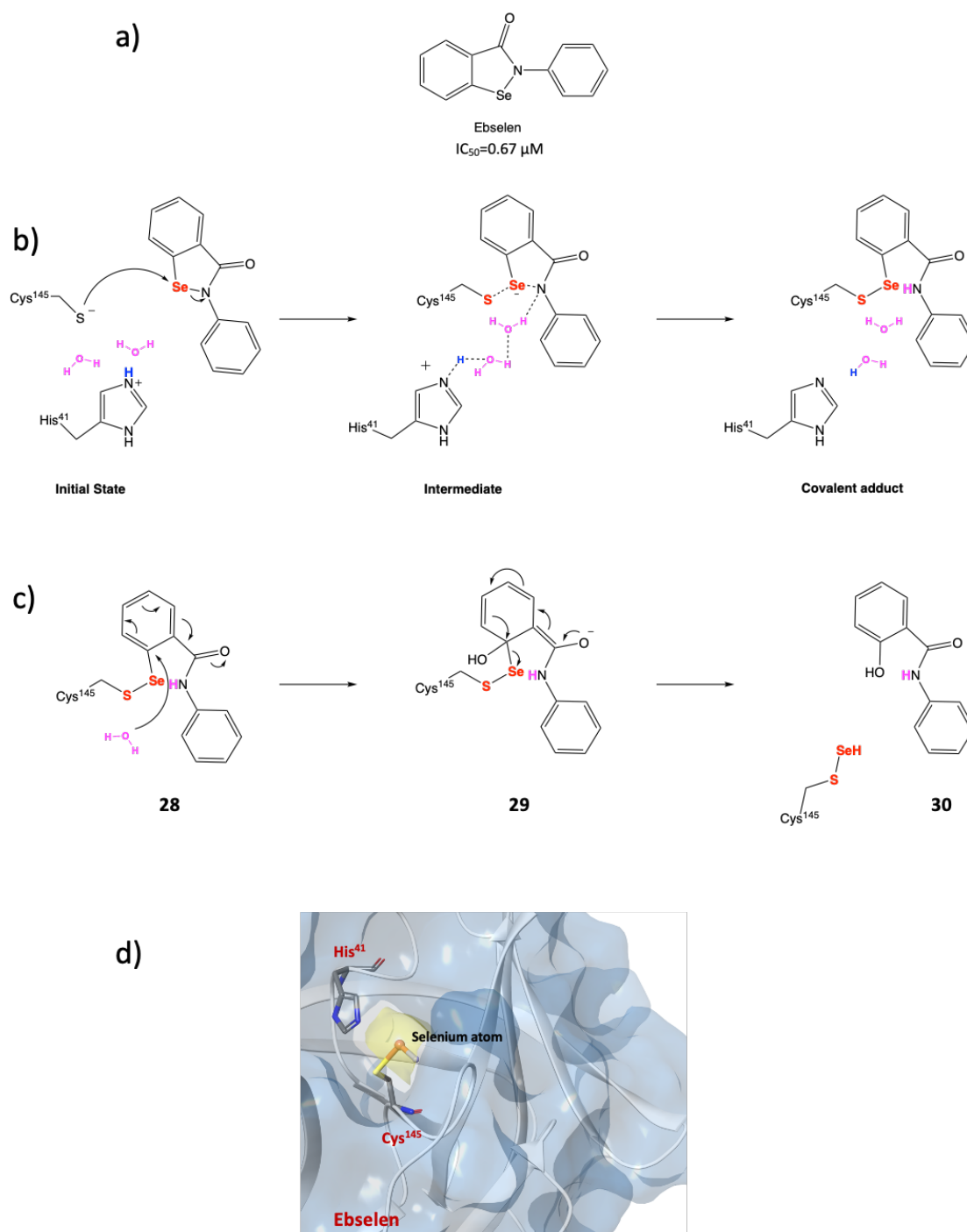
The possibility of using drugs approved in therapy (well-known substances with proved efficacy and safety in humans) in the treatment of COVID-19 allows to reduce the time and costs associated with the development of a new molecule.

*Ebselen* (2-phenyl-1,2-benzisoseleazol-3(2H)-one, in Figure 48a), a heterocyclic structure with a selenium atom that has been studied as an antioxidant/anti-inflammatory agent [185,186], was one of the most interesting examples of repurposing of investigational drugs against SARS-CoV-2.

Thank to *in silico* and *in vitro* studies, *ebsele* demonstrated potent antiviral activity (IC<sub>50</sub>=0.67 μM against SARS-CoV-2 M<sup>PRO</sup>) and capability to covalently bind the catalytic Cys<sup>145</sup> of SARS-CoV-2 M<sup>PRO</sup> [70,187].

To further explore the mechanism of covalent inhibition of this compound, a combined Docking and Density Functional Theory (DFT) approach was used, demonstrating that *ebsele* is able to form a covalent adduct with Cys<sup>145</sup> by the formation of a selenyl sulfide bond [188]. In the Figure 48b it is shown the mechanism of action of *ebsele*: the first step is the activation of the thiol group of Cys<sup>145</sup> through deprotonation mediated by His<sup>41</sup>; subsequently, the activated thiolate performs the nucleophilic attack on the electrophilic selenium atom, determining the opening of the 5-membered ring and the formation of selenyl sulfide bond, responsible for the covalent inhibition of the target. The process is mediated by a molecule of water that acts as a proton transporter [189]. A mass spectrometry study suggested an additional rearrangement of the covalent adduct (Figure 48c), consisting in the hydrolysis of the *ebsele*-SARS-CoV-2 M<sup>PRO</sup> adduct (**28**), by a conserved molecule of water, forming an intermediate state (adduct **29**), with subsequently selenylation of the cysteine and release of a secondary phenolic product **30**. The proposed mechanism of action was confirmed by the

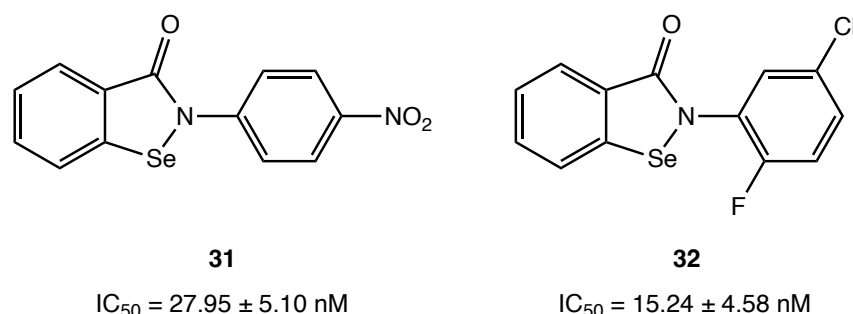
X-ray structure shown in Figure 48d (PDB code: 7BAK), where the selenium atom is covalently bonded to the cysteinyl sulfur [190].



**Figure 48.** (a) Chemical structure of *ebsele*n; (b) proposed mechanism of action of *ebsele*n with SARS-CoV-2 M<sup>PRO</sup>; (c) proposed mechanism for the selenylation of the catalytic site of SARS-CoV-2 M<sup>PRO</sup>; (d) crystal structure of the selenylated cysteine<sup>145</sup> in the SARS-CoV-2 M<sup>PRO</sup> binding site (PDB code: 7BAK) [188-190].

Moreover, *ebsele*n was found to bind an allosteric site, between the I and II domains of M<sup>PRO</sup> (which is essential for the dimerization process) [190,191].

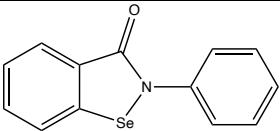
In view of the aforementioned data, *eb-selen* and its analogues could represent promising lead compounds for the future development of new and more effective covalent inhibitors. In this light, many efforts have been made to analyze the SAR of the substituted N-phenyl ring. Indeed, a collection of *eb-selen*-related compounds has been extensively studied. Among them, **31** and **32** (4-nitro and 5-chloro-2-fluoro derivatives, respectively; Figure 49), displayed inhibitory activity superior to that of *eb-selen*, with IC<sub>50</sub> values of 27.95 ± 5.10 nM and 15.24 ± 4.58 nM, respectively [192]. According to SAR analysis, the authors demonstrated that the insertion of one or two substituents on the phenyl ring can improve inhibition activity against SARS-CoV-2 M<sup>PRO</sup> [192].

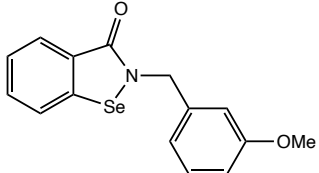
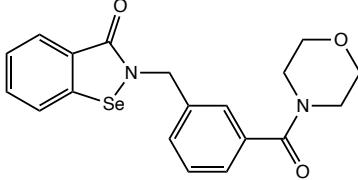
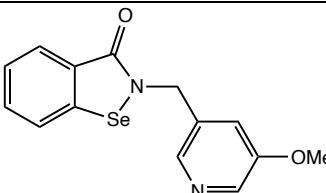
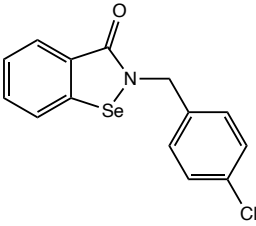


**Figure 49.** *Ebselen* derivatives **31** and **32** [192].

Similarly, four *eb-selen*-based compounds (**33-36**, see Table 13) exhibited greater effectiveness than *eb-selen* both against M<sup>PRO</sup> (IC<sub>50</sub> in the sub-micromolar range) as well as in the viral cell model of SARS-CoV-2 replication [190].

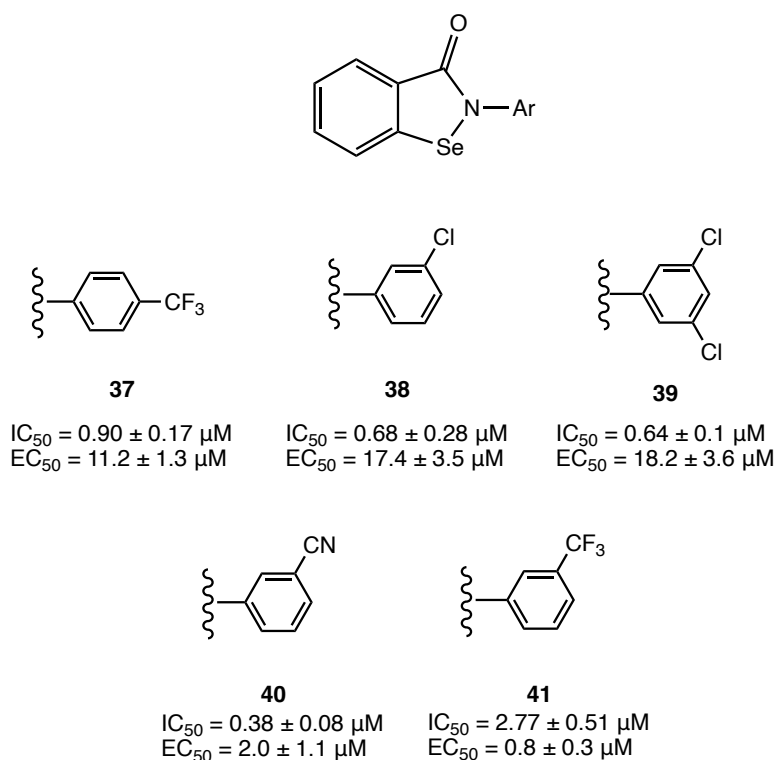
**Table 13.** Chemical structures, SARS-CoV-2 M<sup>PRO</sup> inhibitory activity and inhibition of viral replication data of *eb-selen* derivatives **33-36** [190].

Compound	SARS-CoV 2 M <sup>PRO</sup> inhibition (IC <sub>50</sub> )	Viral replication in Vero E6 cells (EC <sub>50</sub> )
 <i>Ebselen</i>	0.670 μM	4.67 μM

 <p style="text-align: center;"><b>33</b></p>	0.363 $\mu\text{M}$	4.50 $\mu\text{M}$
 <p style="text-align: center;"><b>34</b></p>	0.345 $\mu\text{M}$	3.74 $\mu\text{M}$
 <p style="text-align: center;"><b>35</b></p>	0.467 $\mu\text{M}$	3.17 $\mu\text{M}$
 <p style="text-align: center;"><b>36</b></p>	0.824 $\mu\text{M}$	1.78 $\mu\text{M}$

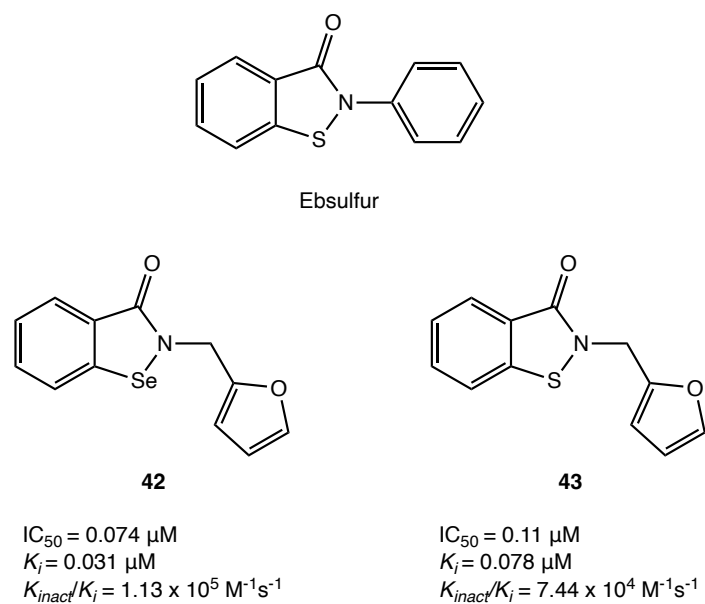
Compounds **37-41** (Figure 50) exhibited sub-micromolar  $\text{IC}_{50}$  values against SARS-COV-2  $\text{M}^{\text{PRO}}$  (in the range 0.38-2.77  $\mu\text{M}$ ), suggesting that the substitution is favorable in *meta* position. As for derivative **40**, the most potent compound, the presence of a cyano group in *meta* improve the interaction network with additional favorable hydrogen bond. Nevertheless, biological assays conducted in an *in vitro* cellular model of SARS-CoV-2 replication (Vero E6 infected cells), showed that, compound **41**, which was the least active in the enzymatic inhibition assay, was surprisingly the most effective in blocking viral progression ( $\text{EC}_{50} = 0.8 \mu\text{M}$ ). For this reason it was selected for further studies in lung organoids, which confirmed the lack of toxicity and its ability to block viral replication [193].





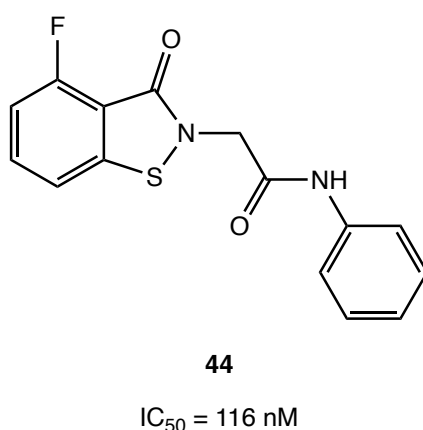
**Figure 50.** Chemical structures of phenyl substituted *ebselen* derivatives **37-41** [193].

From a different point of view, the isosteric analogue of *ebselen*, *ebsulfur* (the selenium atom is replaced by a sulphur atom; Figure 51) has been investigated. Docking studies have shown that *ebsulfur*, in the same way as *ebselen*, can form an S-S covalent bond with the cysteinyl -SH of the SARS-CoV-2 M<sup>PRO</sup>. From this assumption, they designed and biologically evaluated different *ebselen/ebsulfur* analogues as SARS-CoV-2 M<sup>PRO</sup> inhibitors, and, among all, **42** and **43** were the most interesting, with  $IC_{50}$  values of 0.074 and 0.11  $\mu\text{M}$ , respectively (Figure 51). Docking studies proved that the furan group is essential to form additional hydrophobic interactions with Met<sup>165</sup>, Arg<sup>188</sup>, Asp<sup>187</sup>, and Met<sup>49</sup>. From this study it emerged that *ebsulfur*, as well as *ebselen*, could be a potential lead compound for the development of novel, broad spectrum anti coronaviral drugs [194].



**Figure 51.** Chemical structures of *ebsulfur* and *ebselen/ebsulfur* analogues **42** and **43** [194].

More recently, the structure-activity relationship of *ebsulfur* focused on three components: the phenyl ring, the linker and the benzothiazolone core. Compound **44** (Figure 52) as the best covalent inhibitor of SARS-CoV-2 M<sup>PRO</sup>. The phenyl and the benzo-isothiazolone rings were retained in view of their importance in the interaction with the target protein; instead, the introduction of an acetamide group in the linker had a positive effect on the activity, which was confirmed by the remarkable  $IC_{50}$  value of 116 nM against SARS-CoV-2 M<sup>PRO</sup> [195].



**Figure 52.** Chemical structure of *ebsulfur* analogue **44** [195].

### 3.2.1.7. Electrophilic Warheads Reactivity

Binding affinity between the inhibitor and enzyme is crucial for understanding the formation of the enzyme-inhibitor covalent complex. Assessing the binding characteristics of a

compound is essential for predicting and evaluating potential drug interactions. Moreover, an important factor to consider when developing a therapeutic protease inhibitor is the reversibility of compound binding. Irreversible protease inhibitors can produce long-lasting effects by permanently blocking proteases and covalently modifying proteins, which can be beneficial in therapeutic contexts. This understanding aids in the design of effective inhibitors that can selectively target proteases involved in viral replication.

$\alpha$ -ketoamide and carbonyl groups can act either reversible or irreversible derivatives. For example, enzymatic kinetic studies have shown that compounds UAWJ246 and UAWJ247 bind to M<sup>PRO</sup> reversibly, with inhibition constant  $k_i$  values of  $0.036 \pm 0.007 \mu\text{M}$  and  $0.035 \pm 0.008 \mu\text{M}$ , respectively. In contrast, compounds UAWJ248 and GC-376 exhibit an initial reversible binding following by irreversible inactivation, with  $K_{inact}/K_i$  values of  $9.04 \times 10^4 \text{ M}^{-1}\text{s}^{-1}$  and  $2.84 \times 10^4 \text{ M}^{-1}\text{s}^{-1}$ , respectively, suggesting greater potency for UAWJ248 [133].

Michael acceptor derivatives are recognized as irreversible SARS-CoV-2 M<sup>PRO</sup> inhibitors. Kinetics characterization of compounds N3 and myricetin demonstrated an irreversible two-step inhibition of the enzyme [13,70,157,164]. On the other hand, indole core esters serve as irreversible inhibitors of M<sup>PRO</sup>, with compounds GRL-1720 and GRL-0820 showing  $k_i$  values of  $2.15 \mu\text{M}$  and  $0.073 \mu\text{M}$ , respectively, exhibiting strong M<sup>PRO</sup> inhibitory activity. Their second-order rate constants of inactivation demonstrated a time-dependent inhibition, indicating that these compounds act through acylation of the active-site cysteine [147,183,196].

Finally, *ebselen* and *ebsulfur* derivatives exhibited a concentration- and time-dependent inhibition pattern against M<sup>PRO</sup>, with a biphasic character, indicating the rate of inactivation follows pseudo-first-order rate kinetics, which implied that irreversibly covalent inhibition [194].

Following the thorough analysis of the findings presented here, it has become increasingly clear that both non-peptidomimetic and peptidomimetic compounds play a crucial role in enhancing therapeutic properties through their unique chemical characteristics. Building upon this understanding, my research focused on the computational identification and synthesis of compounds possessing a highly reactive electrophilic warhead.

Specifically, in this stage of my research, I focused my attention on the evaluation of various ester compounds as non-peptidomimetic inhibitors, alongside aldehyde compounds as peptidomimetic inhibitors.

### 3.2.2. Covalent inhibition strategy through non-peptidomimetic inhibitors:

#### Synthesis and *in vitro* exploration

Ester-based covalent inhibitors hold significant promise as starting points for Fragment-Based Drug Discovery (FBDD). FBDD focuses on identifying small, simple molecular fragments that, despite their low molecular weight and initial affinity, can specifically bind to biological targets. Once such a fragment is identified, it can be optimized using techniques like fragment growth or linking to create compounds with improved affinity and specificity [197-199]. The initial small size of ester-based inhibitors makes them particularly suitable as starting fragments in FBDD, facilitating the exploration of diverse chemical spaces and the development of potent antiviral agents against challenging targets like SARS-CoV-2 M<sup>PRO</sup> [200]. In this section, we synthesized a series of 6-chloropyridinyl ester compounds, specifically designed as ester-based covalent inhibitors targeting the SARS-CoV-2 M<sup>PRO</sup>. These inhibitors exploit the electrophilic nature of the 6-chloropyridinyl ester warhead to form a covalent bond with the catalytic cysteine residue (Cys<sup>145</sup>) within the active site of M<sup>PRO</sup>. This covalent interaction is crucial for the irreversible enzyme inhibition, which is essential for halting viral replication. Enzymatic inhibition assays confirmed the ability of these synthesized compounds to effectively interact with SARS-CoV-2 M<sup>PRO</sup>. Although these 6-chloropyridinyl esters have demonstrated promising inhibitory potency, they also offer a valuable starting point for further optimization through FBDD.

#### 3.2.2.1. Database design and synthesis

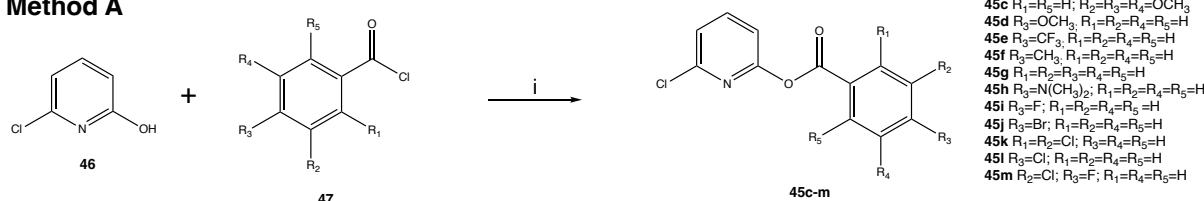
The set of 6-chloropyridinyl ester fragments **45a-m** was synthesized following the two one-pot procedures reported in Scheme 1. In detail, following Method A, the activated ester moiety was readily afforded through the condensation between 2-chloro-6-hydroxypyridine **46** and variously substituted commercial benzoyl chlorides **47**, which provide a highly reactive fragment even towards the mild nucleophilic phenol group of **48**. The use of pyridine as basic catalyst and of anhydrous DCM as solvent provided ideal esterification conditions, allowing to afford the title products in a relatively short reaction time and at RT.

According to the second procedure (Method B in Scheme 1), the combination of carboxyl activating/dehydrating reagent dicyclohexylcarbodiimide (DCC) and of 4-dimethylaminopyridine (DMAP, basic catalyst) provided a useful method for *in situ* activation

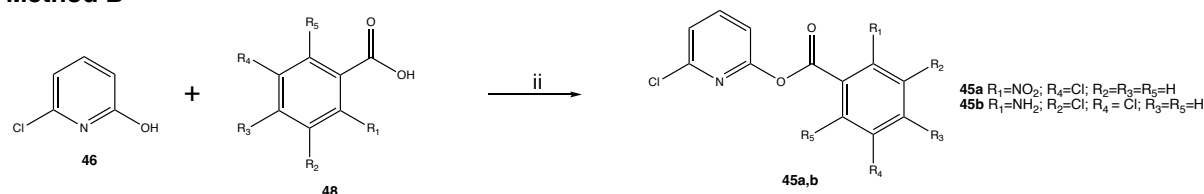
of variously commercial benzoic acids for reaction with the 2-chloro-6-hydroxypyridine **46**. The use of strictly anhydrous conditions and of a nitrogen atmosphere were required to afford the desired products at RT.

For both methods, the resulting esters were purified by silica gel chromatography using a mixture 10:1 petroleum ether/EtOAc in gradient as eluant, providing sufficiently air/moisture-stable ester derivatives in good to excellent yields (70-95%).

#### Method A



#### Method B



**Scheme 1.** Synthesis of 6-chloropyridinyl ester derivatives **45a-m**. Reagents and conditions: (i): 2-Chloro-6-Hydroxy-pyridine (1 eq.), appropriate benzoyl chloride (1.2 eq.), pyridine (1.2 eq.), anhydrous DCM, rt 12-24h; (ii): 2-Chloro-6-Hydroxy-pyridine (2 eq.), appropriate carboxylic acid (2 eq.), DCC (3 eq.), DMAP (1 eq.), N<sub>2</sub>, anhydrous DCM, rt 12-24 h.

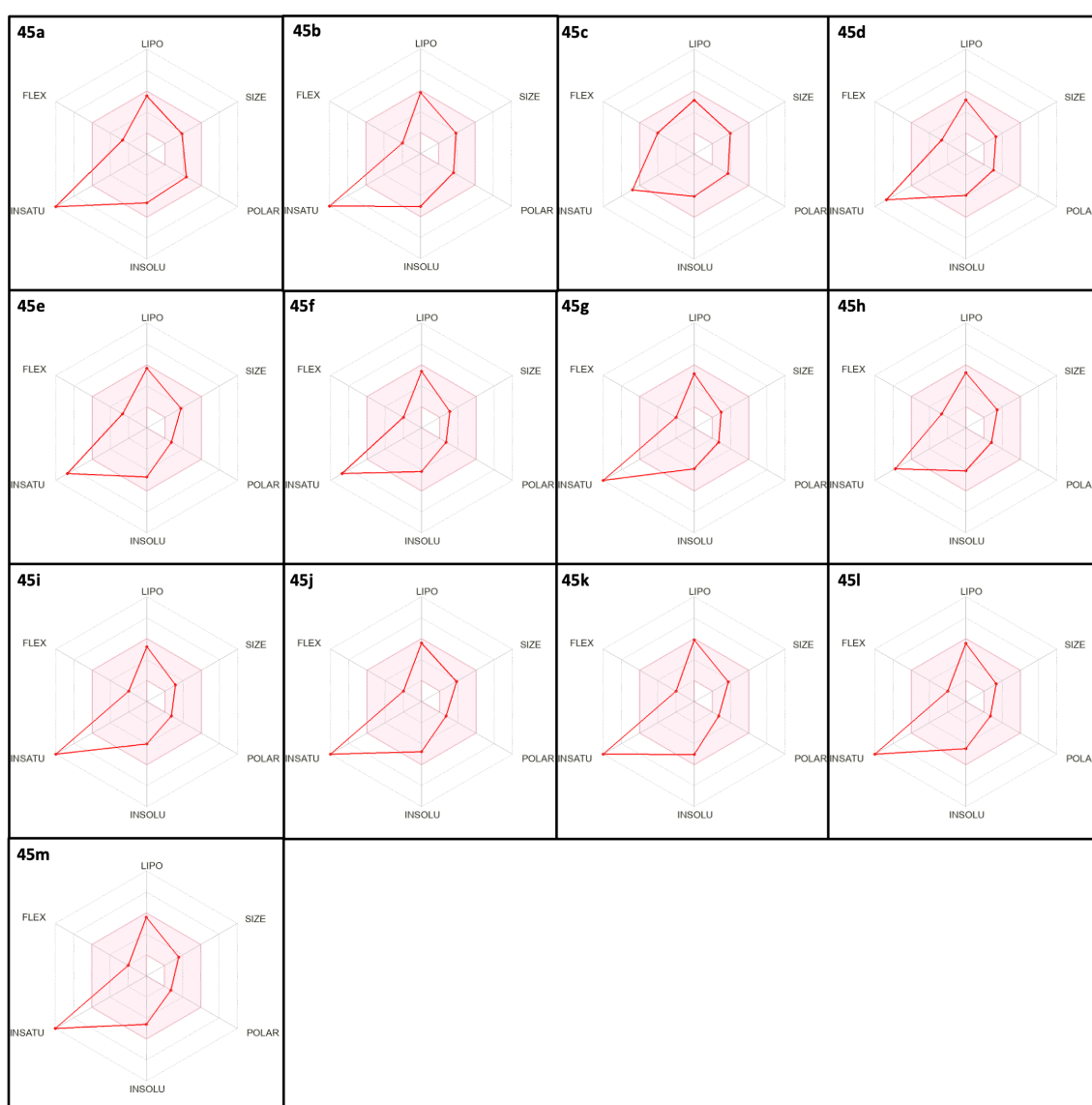
### 3.2.2.2. ADME properties

In drug discovery, the preliminary estimation of ADME (Absorption, Distribution, Metabolism, and Excretion) and drug-likeness parameters using *in silico* techniques is an invaluable aid to save both time and resources, increase the success rate, and consequently reduce the risks of failure in the preclinical and clinical phases [201].

To gain insight into the drug-likeness of our compounds, we decided to use the SwissADME tool (<http://www.swissadme.ch>) [111], considering a set of well consolidated parameters for searching bioactive compounds, such as PAINS filters [112], Lipinski's rules [113], Veber [114], and Egan filters [115]. The analysis of the data highlighted that the ester derivatives, as small molecules with low molecular weight (<350), met the expectations in terms of bioactivity. Indeed, all compounds have no violations and PAINS.

In addition, a bioavailability radar plot was created for each compound, allowing a rapid and graphical appraisal of drug-likeness, Figure 53. The hexagonal graph shows six axes

representing six important properties for oral bioavailability: lipophilicity, size, polarity, solubility, fraction of saturated bonds, and flexibility. The range of optimal values for each parameter defines the pink area in the plot, and the radar plot for a molecule should fall almost entirely within this colored area (for further details about the parameters considered, see reference [111]). Except for the parameters fraction of saturated bonds and flexibility, which are lower than the standard data, the values of the lipophilicity, size, polarity, solubility parameters are in the optimal range, Figure 53. See Supplementary Material S1, Matrix S2 for the complete output matrix from SwissADME.



**Figure 53.** Bioavailability radar plot produced using the SwissADME tool calculations for compounds 45a-m.

All compounds were predicted to be highly or completely absorbed in the gastrointestinal tract through a passive mechanism. Furthermore, esters exhibited an optimal balance between lipophilicity and hydrophilicity. Derivatives **45a-m** do not appear to be substrates of P-gp, one of the best characterized efflux pumps that frequently contributes to suboptimal bioavailability and drug resistance in cancer cells (active efflux mechanisms). LogP, water solubility, and P-gp parameters are listed in the Supplementary Material S1, Matrix S2.

### 3.2.2.3. Enzyme inhibition assay of SARS-CoV-2 M<sup>PRO</sup>

The activity of M<sup>PRO</sup> was assessed using the SensoLyte 520 SARS-CoV-2 3CL<sup>PRO</sup> Activity Assay Kit. Compounds were serially diluted from 5000  $\mu$ M to 1  $\mu$ M and tested for their ability to inhibit the protease activity. The assay results revealed that compound **45g** exhibited the lowest IC<sub>50</sub> value of  $30 \pm 6.6$   $\mu$ M, indicating a significant inhibitory effect and positioning it as a promising candidate for further development as an antiviral agent. Conversely, the IC<sub>50</sub> values for compounds **45c**, **45d**, **45e**, **45h**, and **45j** could not be calculated, as these compounds demonstrated minimal inhibition, with percentages close to zero even at the highest tested concentration of 500  $\mu$ M. This suggests that these compounds may lack sufficient affinity for the 3CL<sup>PRO</sup> target or may require structural optimization to enhance their inhibitory efficacy. The identification of **45g** as a potent inhibitor underscores the potential of targeting the M<sup>PRO</sup> enzyme in the development of novel therapeutic strategies against COVID-19. The results, including all IC<sub>50</sub> values, are detailed in Table 14, highlighting the need for continued research into the structure-activity relationships of these compounds to refine their pharmacological properties and enhance their effectiveness in viral inhibition.

**Table 14.** IC<sub>50</sub> Values of Compounds Tested for Inhibition of SARS-CoV-2 M<sup>PRO</sup>.

Compound	IC <sub>50</sub> ( $\mu$ M)
<b>45a</b>	195,8 $\pm$ 65
<b>45b</b>	>500
<b>45c</b>	>500
<b>45d</b>	>500
<b>45e</b>	>500
<b>45f</b>	115,9 $\pm$ 40
<b>45g</b>	30 $\pm$ 6,6

<b>45h</b>	>500
<b>45i</b>	73,4±20
<b>45j</b>	>500
<b>45k</b>	>500
<b>45l</b>	>500
<b>45m</b>	>500

#### 3.2.2.4. Future Perspectives

The synthesized compounds, particularly the most promising candidate, **45g**, will undergo comprehensive further analysis. This process will begin with an assessment of their ability to form covalent bonds, starting with glutathione, a key biomolecule involved in cellular redox balance and detoxification. This initial evaluation aims both the reactivity of the compounds and their capacity to engage in covalent interactions, which is essential for their potential as irreversible inhibitors. Following this, we will investigate the compounds capacity to form a covalent adduct directly with a cysteine residue, mimicking the catalytic Cys<sup>145</sup> of SARS-CoV-2 M<sup>PRO</sup>. This step is critical, as it will provide insight into the selectivity and stability of the covalent bond, which directly influences the effectiveness of the inhibitors. Confirming these interactions will establish a solid foundation for the compound's development as therapeutic agents. Once the covalent interactions are validated, the compounds will proceed to a rigorous Lead Optimization process, leveraging their profiles as starting fragments for FBDD. This stage will involve detailed *in silico* evaluations, including molecular docking studies to predict binding affinities and orientations, such as IFD to account for conformational changes upon binding, and Covalent Docking (CovDock) to specifically model the formation of covalent adducts. Additionally, Molecular Dynamics Simulations will be utilized to assess the stability of the inhibitor-enzyme complexes over time, allowing for a deeper understanding of the dynamic behavior of the compounds within the active site of M<sup>PRO</sup>. Ultimately, these efforts aim to advance the development of novel therapeutics capable of effectively combating SARS-CoV-2 and other related viral infections.



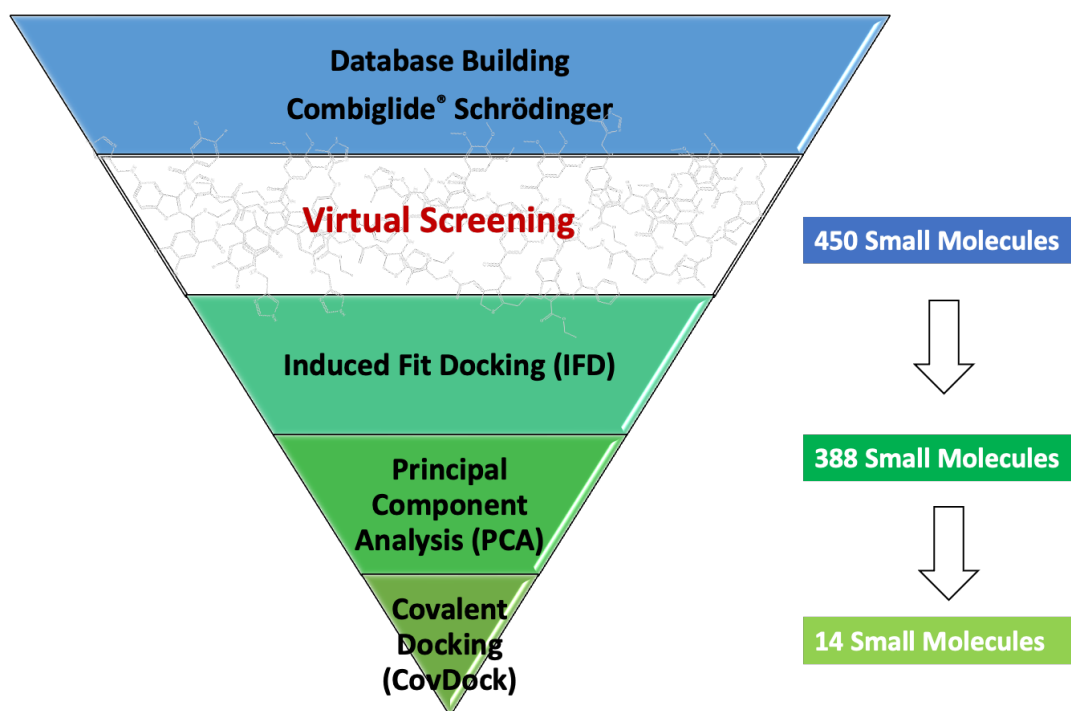
### 3.2.3. Covalent inhibition strategy through peptidomimetic inhibitors: Computational Studies

In response to the growing interest in more potent and effective covalent inhibitors, this step of my research project proposes a rational design of peptidomimetic compounds, with aldehydic warhead, as promising SARS-CoV-2 M<sup>PRO</sup> covalent inhibitors.

Our approach involved a systematic workflow, as shown in Figure 54, structured in key phases to identify potential inhibitors. Initially, we identified the most recurrent and crucial fragments for each portion of the inhibitor (P1, P2, P3), as outlined in section 3.2.1 [76]. Using a combinatorial approach (CombiGlide) integrated into the Maestro software suite, we constructed a database of 450 molecules, each designed with an aldehydic warhead at the P1' position to facilitate potential covalent interaction with the active site cysteine of M<sup>PRO</sup>.

Subsequently, IFD studies were performed on these molecules within the M<sup>PRO</sup> binding site. Out of the 450 molecules, 388 were successfully docked, while 62 were excluded due to size incompatibility with the binding site. For the docked structures and a reference set of known M<sup>PRO</sup> inhibitors with established IC<sub>50</sub> values, we used DRUDIT and MOLDESTO to calculate molecular descriptors. Based on these descriptors, we developed a PCA model, which produced a 2D plot where inhibitors with IC<sub>50</sub> values below 0.1 μM and those above 0.1 μM formed two distinct clusters. We then analyzed the distribution of the 388 molecules within the PCA plot to assess their positioning relative to the two clusters, providing insights into their potential inhibitory efficacy.

Based on their favorable positioning in the PCA plot and alignment with the identified structural requirements, we selected 14 molecules for detailed CovDock studies.



**Figure 54.** Virtual Screening workflow, including database building, and the three key steps (IFD, PCA, and CovDock), for the identification of potential peptidomimetic covalent inhibitors with aldehydic warhead.

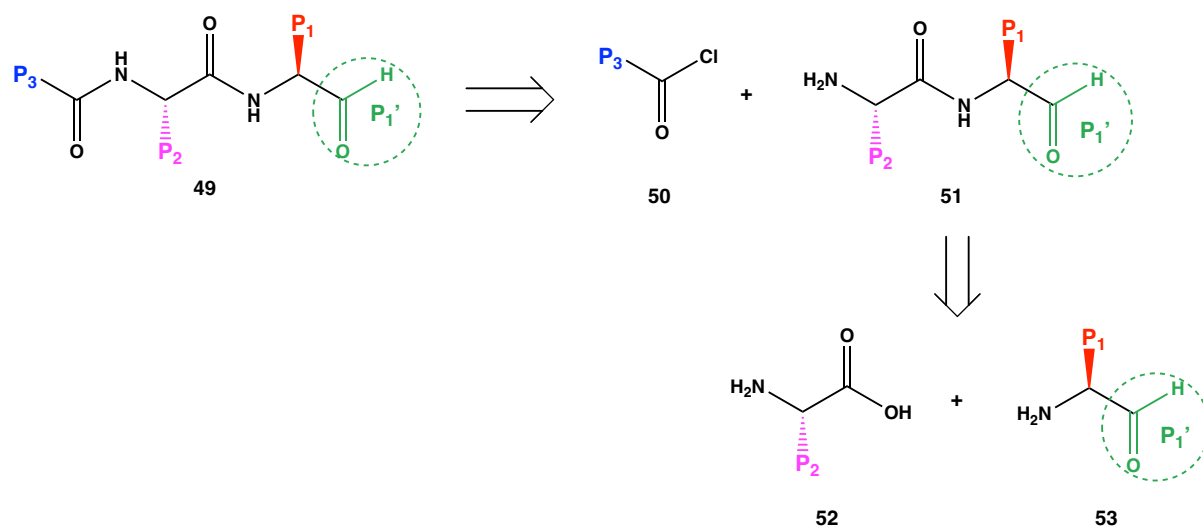
Through this structured and methodical approach, we successfully identified promising new covalent peptidomimetic inhibitors featuring aldehydic warheads that demonstrate strong potential as SARS-CoV-2 M<sup>PRO</sup> inhibitors. These findings represent a significant contribution to the ongoing global effort in antiviral drug development and offer a robust framework for the future design of selective covalent inhibitors.

### 3.2.3.1. *In silico* design of a new database of small molecules using CombiGlide® Schrödinger

The initial phase of the research project focused on constructing a comprehensive database of peptidomimetic structures, each containing a reactive electrophilic aldehyde warhead. This was accomplished using CombiGlide®, a tool within the Maestro software suite. The database construction process consisted of three sequential steps: designing a retrosynthetic scheme, preparing the reagents, and constructing the combinatorial library.

To efficiently plan the computational synthesis of aldehyde-based peptidomimetic derivatives **49**, a well-structured retrosynthetic scheme was developed. As outlined in Scheme 2, the synthesis of these derivatives follows a pathway designed solely for computational purposes;

therefore, specific reaction conditions and experimental details are not included in this analysis. The computational scheme is developed to virtually generate the molecular files corresponding to derivatives **49**. Initially, the target molecules are formed via a nucleophilic acyl substitution reaction between synthons **50** and **51**, establishing the crucial acyl linkage that characterizes the peptidomimetic structure. Prior to this, synthon **51** is synthesized through a peptide coupling reaction between synthons **52** and **53**, assembling the necessary peptidic backbone. This retrosynthetic strategy ensures a systematic and coherent approach to the construction of the desired peptidomimetic derivatives, integrating all essential structural features to achieve the targeted biological activity.



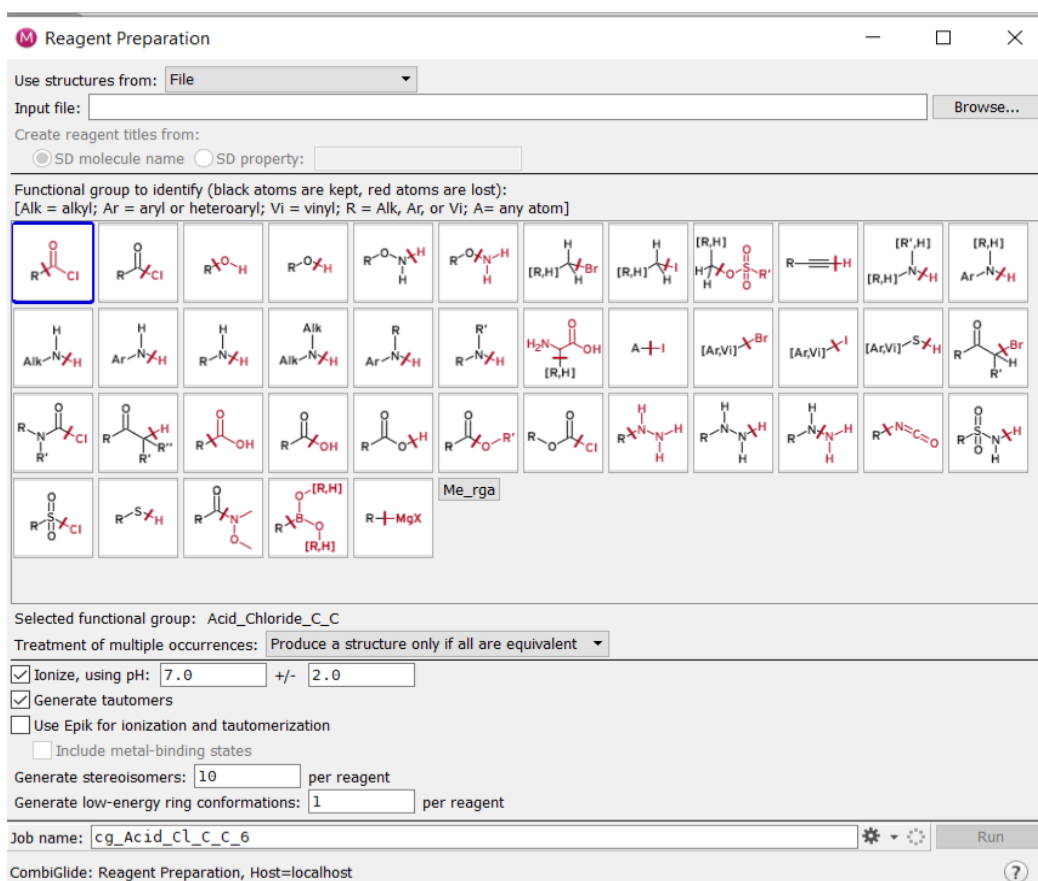
**Scheme 2.** Retrosynthetic scheme of aldehyde-based peptidomimetic derivatives **49**. The electrophilic warhead in  $P_1'$  is shown in green, while the  $P_1$ – $P_3$  fragments are depicted in red, purple, and blue, respectively.

Following the retrosynthetic analysis, the second step involves the utilization of the Reagent Preparation Panel (Figure 55). This crucial step ensures that the input files for the synthons (or reagents) are adequately prepared for use in the subsequent combinatorial library enumeration process. The Reagent Preparation Panel is designed to convert 2D molecular structures into optimized 3D structures, embedding all necessary information required to construct molecules for virtual screening.

The preparation begins with the conversion of 2D chemical structures into their corresponding 3D forms, performed by the Ligand Preparation task (LigPrep), which not only generates 3D conformers but also performs energy minimization to ensure that the structures

are in a low-energy, physically realistic state. This step is vital as it reduces the likelihood of artifacts in the modeling process and ensures that the reagents are ready for subsequent computational operations.

The main task of the Reagent Preparation Panel is the selection of specific reagent types based on functional groups. For each reagent, the panel identifies the functional group within the molecule and pinpoints the specific bond that will be replaced or modified when the reagent is added to the core structure during the library enumeration process. This targeted approach ensures that the chemical transformations modeled *in silico* are both realistic and aligned with the intended synthetic strategy.



**Figure 55.** Reagent Preparation Panel (*Maestro Schrödinger*).

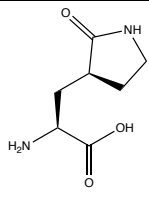
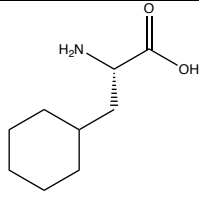
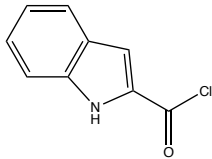
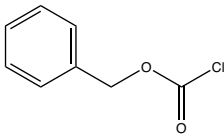
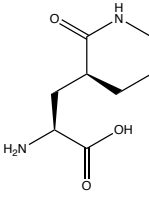
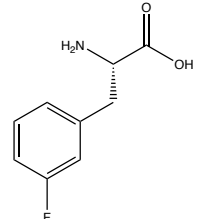
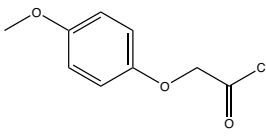
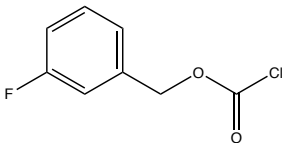
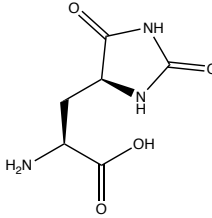
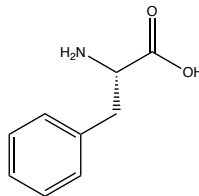
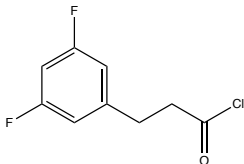
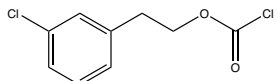
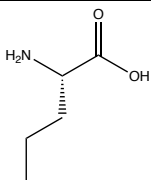
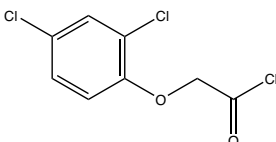
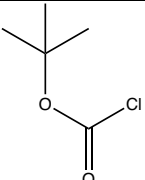
The output of the Reagent Preparation Panel consists of prepared reagent files, ready for integration into the combinatorial library. These files contain 3D structures of the reagents, along with information about the reactive sites. Notably, the process may yield multiple output structures for a single input reagent, reflecting the different energetically favorable conformations that the molecule can adopt. This conformational diversity is crucial for

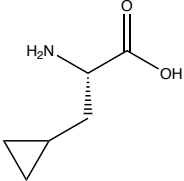
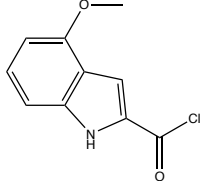
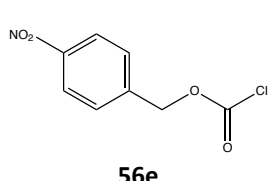
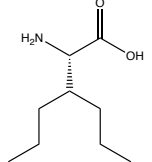
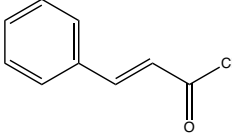
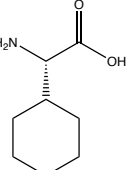
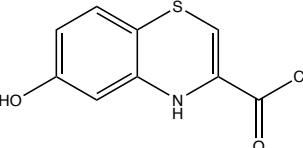
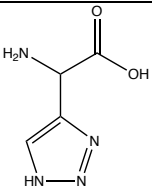
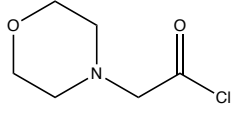
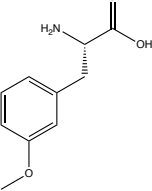
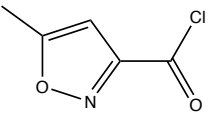
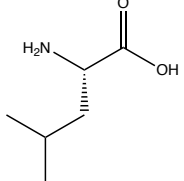
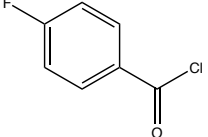
exploring a wide chemical space during virtual screening, as different conformers might interact differently with potential target sites.

This preparation process must be repeated for each type of reagent intended for use in the combinatorial library enumeration. By carefully preparing each reagent, the method ensures that all components of the combinatorial library are structurally sound and ready for subsequent steps in the computational database design workflow, ultimately enhancing the reliability and success of the virtual screening efforts.

Table 15 lists all the building blocks used in the combinatorial process. Specifically, three types of reagents were prepared: amino acids **54a-m** for including P1 and P2 fragments, acyl chlorides **55a-j**, and carbonyl chlorides **56a-e**.

**Table 15.** List of building blocks used in the combinatorial synthesis process. This includes amino acids (**54a-m**) used for P1 and P2 fragments, acyl chlorides (**55a-j**), and carbonyl chlorides (**56a-e**) employed for P3 fragments.

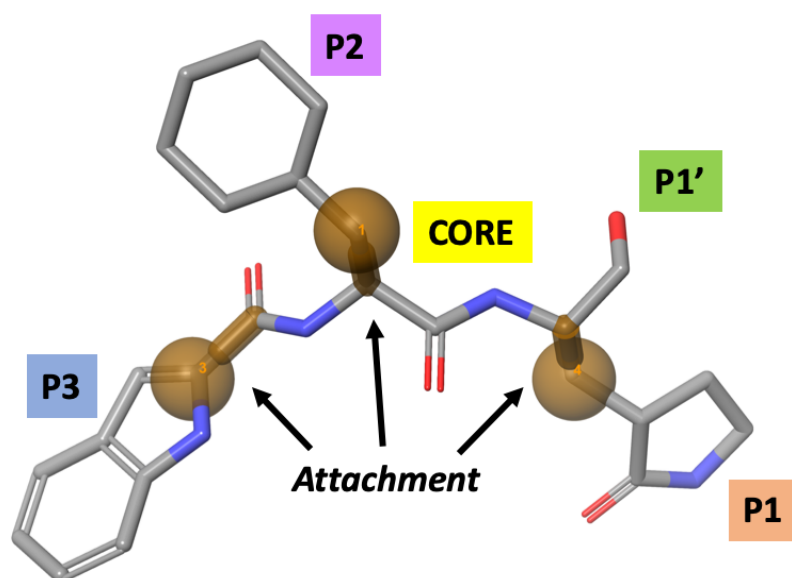
P1	P2	P3	P3
 <b>54a</b>	 <b>54d</b>	 <b>55a</b>	 <b>56a</b>
 <b>54b</b>	 <b>54e</b>	 <b>55b</b>	 <b>56b</b>
 <b>54c</b>	 <b>54f</b>	 <b>55c</b>	 <b>56c</b>
	 <b>54g</b>	 <b>55d</b>	 <b>56d</b>

	 <p><b>54h</b></p>	 <p><b>55e</b></p>	 <p><b>56e</b></p>
	 <p><b>54i</b></p>	 <p><b>55f</b></p>	
	 <p><b>54j</b></p>	 <p><b>55g</b></p>	
	 <p><b>54k</b></p>	 <p><b>55h</b></p>	
	 <p><b>54l</b></p>	 <p><b>55i</b></p>	
	 <p><b>54m</b></p>	 <p><b>55j</b></p>	

Additionally, a core structure was prepared as an example molecule for the combinatorial process.

Once all the reagents were prepared, the third and final step in the library design was carried out using the Combinatorial Library Enumeration Panel to construct a library of compounds for subsequent virtual screening. This panel provides tools to develop a ligand library starting from a core structure to which molecular fragments are added at user-specified positions called “attachments”. Specifically, combinatorial library enumeration is performed by substituting fragments in the attachments of the core structure with building blocks from the

reagent structures. Both the core and the reagents must have designated bonds to break (example in Figure 56). Reagent fragments are then attached to the core at the site of the broken bond in a process referred to as "growing" and the bond being replaced is known as the "grow bond". These bonds are not necessarily the bonds broken and formed in the actual chemical reaction but serve to vary the substituents on a central structure.



**Figure 56.** Example of structure central core with the definition of the 3 attachments required for the process. The electrophilic warhead presents in P1' is represented in green, while the fragments P1–P3 are shown in red, purple, and blue, respectively.

At the end of this process, we obtained a database of 450 molecules with an aldehyde warhead for virtual screening. Supplementary Material S1, Tables S3 and S4 provide the SMILES representations for the fragments and the final obtained compounds.

### 3.2.3.2. Induced Fit Docking studies

Following the construction of a database comprising 450 peptidomimetic small molecules with an aldehyde warhead, IFD studies were conducted as the first step of virtual screening to assess their potential compatibility with the active site of the SARS-CoV-2 M<sup>PRO</sup>.

To perform this analysis, the crystallographic structure of the SARS-CoV-2 M<sup>PRO</sup> with PDB code 7VH8 [116] was selected. This X-ray structure was chosen for its high resolution and its co-crystallized inhibitor with a peptidomimetic core (*nirmatrelvir*, see Figure 13), making it particularly relevant for the evaluation of similar compounds.

The IFD studies led to the selection of 388 molecules from the initial 450, all of which exhibited IFD scores superior to that of the co-crystallized ligand. Supplementary Material S1, Table S5 shows IFD and docking scores for the selected 388 small molecules and *nirmatrelvir*.

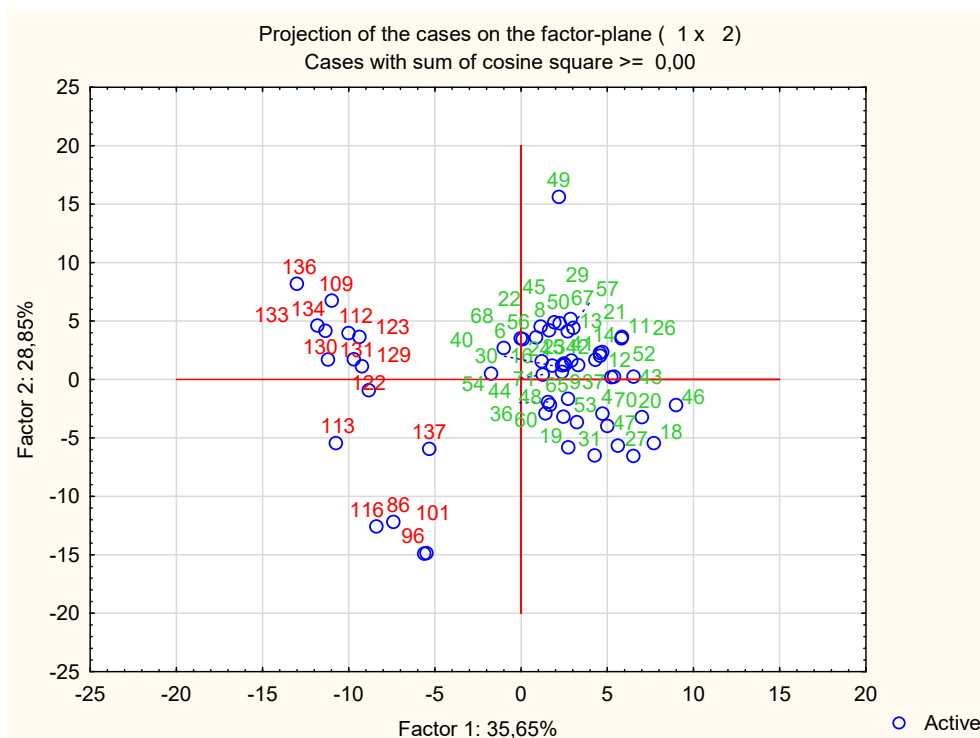
The excluded 62 molecules were not docked by Maestro, because of structural incompatibility with the binding site of SARS-CoV-2 M<sup>PRO</sup>. These molecules, likely due to their larger size or more rigid structural features, could not be successfully docked, suggesting that their steric or conformational properties rendered them unsuitable for fitting within the active site of M<sup>PRO</sup>. This structural incompatibility highlights the importance of molecular size, shape, and flexibility when evaluating compounds for binding potential within a given target site. Consequently, these 62 molecules were excluded from further analysis, allowing the focus to shift to the remaining 388 compounds that exhibited favorable docking results and superior IFD scores compared to the co-crystallized ligand, *nirmatrelvir*.

### 3.2.3.3. DRUDIT and Principal Component Analysis

To further evaluate the 388 molecules selected through IFD, PCA was performed. This analysis involved calculating a set of molecular descriptors for a group of covalent inhibitors, which were previously identified through literature research and for which enzymatic inhibition assays against SARS-CoV-2 M<sup>PRO</sup> had been conducted. Specifically, the calculation of molecular descriptors was performed using MOLDESTO, from DRUDIT (see section 3.1.1.1.) [108]. Supplementary Material S1, Table S6 shows SMILES and IC<sub>50</sub> for structures utilized for the PCA model building.

The application of PCA to the structure *versus* Molecular Descriptors matrix (Supplementary Material S1, Matrix S3) revealed a total variance of 64.5% explained by the first two principal components. The 2D plot (PC1 vs. PC2, Figure 57) highlighted the clustering of the compounds based on their IC<sub>50</sub> values. Specifically, all molecules with an IC<sub>50</sub> < 0.1 μM are represented in green, while all molecules with an IC<sub>50</sub> > 0.1 μM are represented in red.





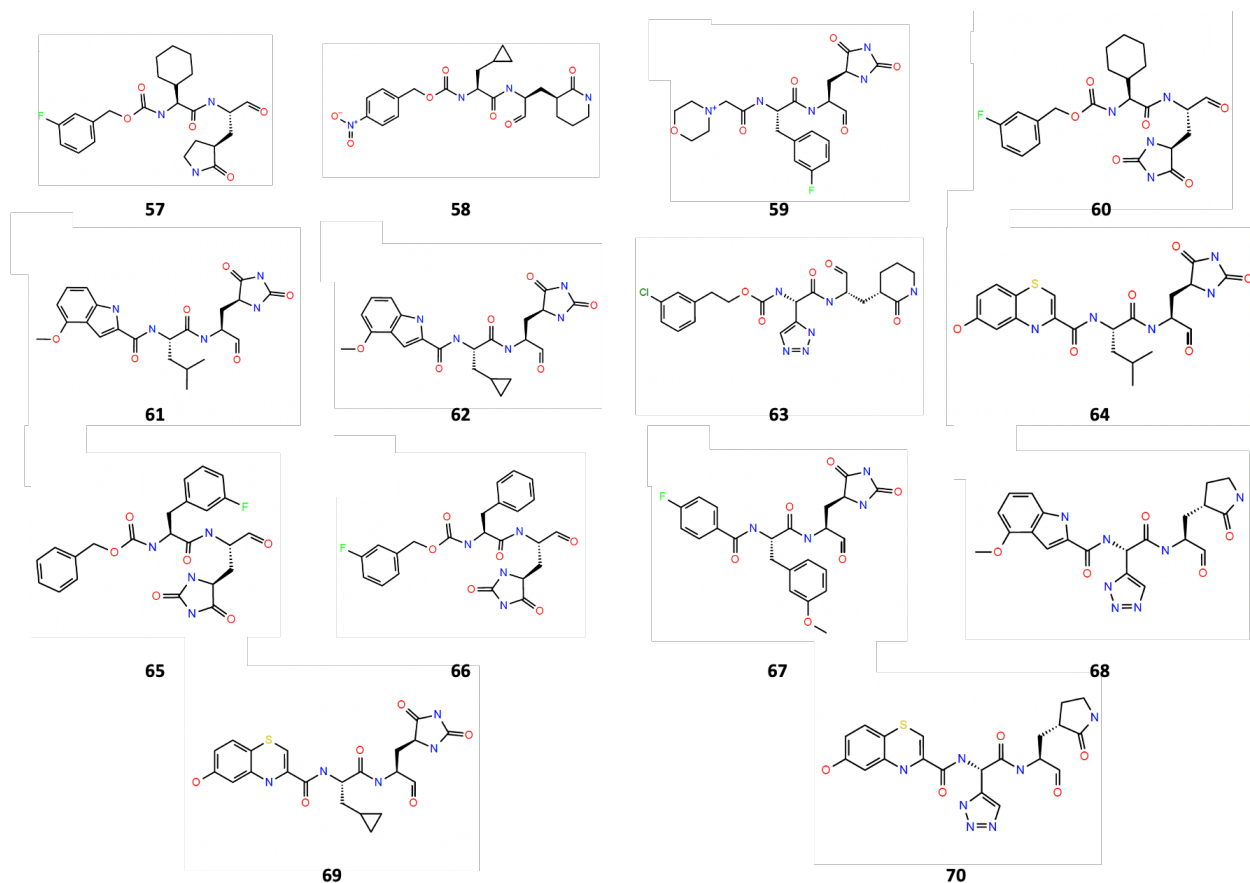
**Figure 57.** Principal Component Analysis (PC1 versus PC2) applied to the molecular descriptors matrix of the selected known SARS-CoV-2 M<sup>PRO</sup> inhibitors.

With the PCA model established, we then assessed the position of each of the 388 molecules in the plot relative to the red or green clusters. This evaluation involved determining whether each molecule fell into the cluster associated with  $IC_{50} < 0.1 \mu\text{M}$  (green) or the cluster with  $IC_{50} > 0.1 \mu\text{M}$  (red). By analyzing these distributions, we aimed to understand how the molecular descriptors influenced the molecules inhibitory activity against SARS-CoV-2 M<sup>PRO</sup>. In details, the selected small molecules were treated as described in the multi-step process below.

1. Calculation of Molecular Descriptors: for each of the 388 molecules, we recalculated the same set of molecular descriptors that were used to build the PCA model matrix (Supplementary Material S1, Matrix S4). This ensured that each molecule was represented consistently in terms of its chemical properties as well as molecules in the known inhibitors set;
2. Centroid Calculation: the centroids, or the central points, of the two clusters identified in the PCA plot (corresponding to the  $IC_{50} < 0.1 \mu\text{M}$  and  $IC_{50} > 0.1 \mu\text{M}$  groups) were calculated (Supplementary Material S1, Table S7). These centroids serve as reference points to evaluate the proximity of each molecule to the clusters;

3. Distance Calculation: for each molecule, we computed the distances from its position in the PCA plot to both centroids (Supplementary Material S1, Table S8). This allowed us to quantify how close each molecule was to the centers of the respective clusters;
4. Distance Ratio: the ratio between the distance from each molecule to the green centroid and the distance to the red centroid was computed (Supplementary Material S1, Table S9). This ratio provides insight into the relative proximity of each molecule to the two clusters: a smaller ratio indicates that the molecule is closer to the green cluster and farther from the red cluster;
5. Cutoff Value: a cutoff value of 0.3 was established for the ratio. Molecules with a ratio less than 0.3 are considered to be closer to the green cluster, indicating they are more similar to the high-potency inhibitors ( $IC_{50} < 0.1 \mu\text{M}$ ) and relatively farther from the lower-potency inhibitors ( $IC_{50} > 0.1 \mu\text{M}$ ).

By following these steps, we were able to systematically evaluate the distribution of each molecule in relation to the defined clusters and assess their potential classification based on their molecular descriptors and PCA positioning. This analysis enabled us to select 14 molecules (derivatives **57-70** in Figure 58) for subsequent Covalent Docking studies.



**Figure 58.** 2D chemical structures of compounds **57-70**, screened through multivariate statistical analysis.

### 3.1.3.4. Covalent Docking Studies

Computational insights into covalent docking are becoming increasingly crucial in addressing the challenges of selectivity and potency in covalent inhibitors. These inhibitors function through a dual mechanism: covalent bond formation between the ligand and the target protein, as well as stabilizing non-covalent interactions within the binding pocket. The CovDock protocol aids in identifying optimal covalent complexes using the well-validated Prime energy model, and calculates an apparent affinity score, which incorporates key elements of an effective covalent docking process.

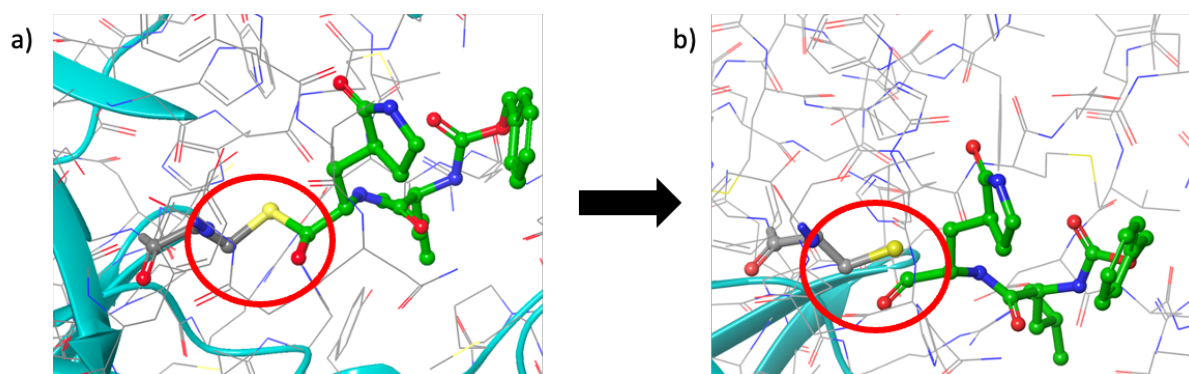
For successful covalent docking, two critical conditions must be met: the pre-reactive form of the ligand must reside in the binding pocket long enough to facilitate the reaction between the ligand warhead and the protein reactive residue, and the docking process must avoid unfavorable steric clashes and poor electrostatic interactions as the reaction progresses.

The CovDock protocol initiates by employing Glide docking on a receptor, temporarily replacing the reactive residue with alanine. Following this, the reactive residue is reinstated

and sampled to form a covalent bond with the ligand, generating multiple binding poses. These covalent complexes are then minimized using the Prime energy model, selecting the top-scoring complexes. Furthermore, an apparent affinity score, derived from the Glide scores of both pre-reactive and post-reactive ligand conformations, is computed to estimate binding energies, providing an effective tool for virtual screening to assess potential inhibitors [202,203].

Compounds **57-70** were submitted to CovDock analysis to evaluate their capacity to covalently bind to the reactive Cys<sup>145</sup> within the SARS-CoV-2 M<sup>PRO</sup> binding site. With this aim, we choose a protein with **GC-373**, see Figure 23, as co-crystallized compound (PDB code 6WTK [126]). In addition to generating the grid, the reaction type was specified by selecting the "Nucleophilic Addition to a Double Bond" option. Furthermore, the covalent bond between **GC-373** and Cys<sup>145</sup> was broken to restore the thiol group (SH) of the cysteine residue, allowing the system to simulate the unbound reactive state of Cys<sup>145</sup> before evaluating the binding potential of the tested compounds.

Figure 59a illustrates the covalent M<sup>PRO</sup>/**GC-373** complex, while Figure 59b displays the broken covalent bond between Cys<sup>145</sup> and **GC-373**, highlighting the restoration of the cysteinyl -SH group.



**Figure 59.** (a) Representation of the covalent complex between SARS-CoV-2 M<sup>PRO</sup> and **GC-373**, highlighting the covalent bond between the Cys<sup>145</sup> residue and the ligand; (b) cleavage of the covalent bond between Cys<sup>145</sup> and **GC-373**, with the restoration of the thiol group (-SH) of the cysteine residue, simulating the unbound reactive state of Cys<sup>145</sup>.

It is important to note that compounds **66** and **70** were excluded during the CovDock calculations performed by Maestro. This exclusion could be attributed to various factors, such as the inability of these compounds to adopt a suitable geometry for covalent bond formation with Cys<sup>145</sup>, or the failure to establish stable non-covalent interactions within the binding

pocket. Additionally, steric clashes or unfavorable electrostatic interactions might have contributed to the software decision to reject these compounds as potential covalent binders. Meanwhile, the docking scores and CovDock affinity values revealed that compound **59** exhibited the highest affinity (-9.609) and the lowest docking score (-10.213), indicating strong potential for covalent binding. Other compounds, such as **62** (-9.431), **61** (-9.230), and **64** (-8.858) also demonstrated notable affinity. In contrast, the reference compound **GC-373** showed a significantly lower affinity (-6.916) and a docking score of -6.335, suggesting that several of the identified compounds may outperform **GC-373** in binding to the target site. These results provide valuable insights into the potential of compounds **57-65** and **67-69** as effective covalent inhibitors. Docking score and CovDock affinity values for compounds **57-65** and **67-69** and the reference **GC-373** are shown in Table 16.

**Table 16.** *CovDock affinity and docking score values for compounds 57-65, 67-69, and GC-373.*

<b>Compound</b>	<b>CovDock Affinity</b>	<b>Docking Score</b>
<b>59</b>	-9.609	-10.213
<b>62</b>	-9.431	-9.531
<b>61</b>	-9.230	-8.970
<b>64</b>	-8.858	-9.746
<b>60</b>	-8.849	-9.770
<b>58</b>	-8.608	-9.566
<b>57</b>	-8.520	-9.333
<b>69</b>	-8.454	-8.727
<b>68</b>	-8.156	-8.016
<b>67</b>	-7.933	-7.905
<b>GC-373</b>	-6.916	-6.335
<b>65</b>	-6.211	-6.879

<b>63</b>	-6.167	-6.892
-----------	--------	--------

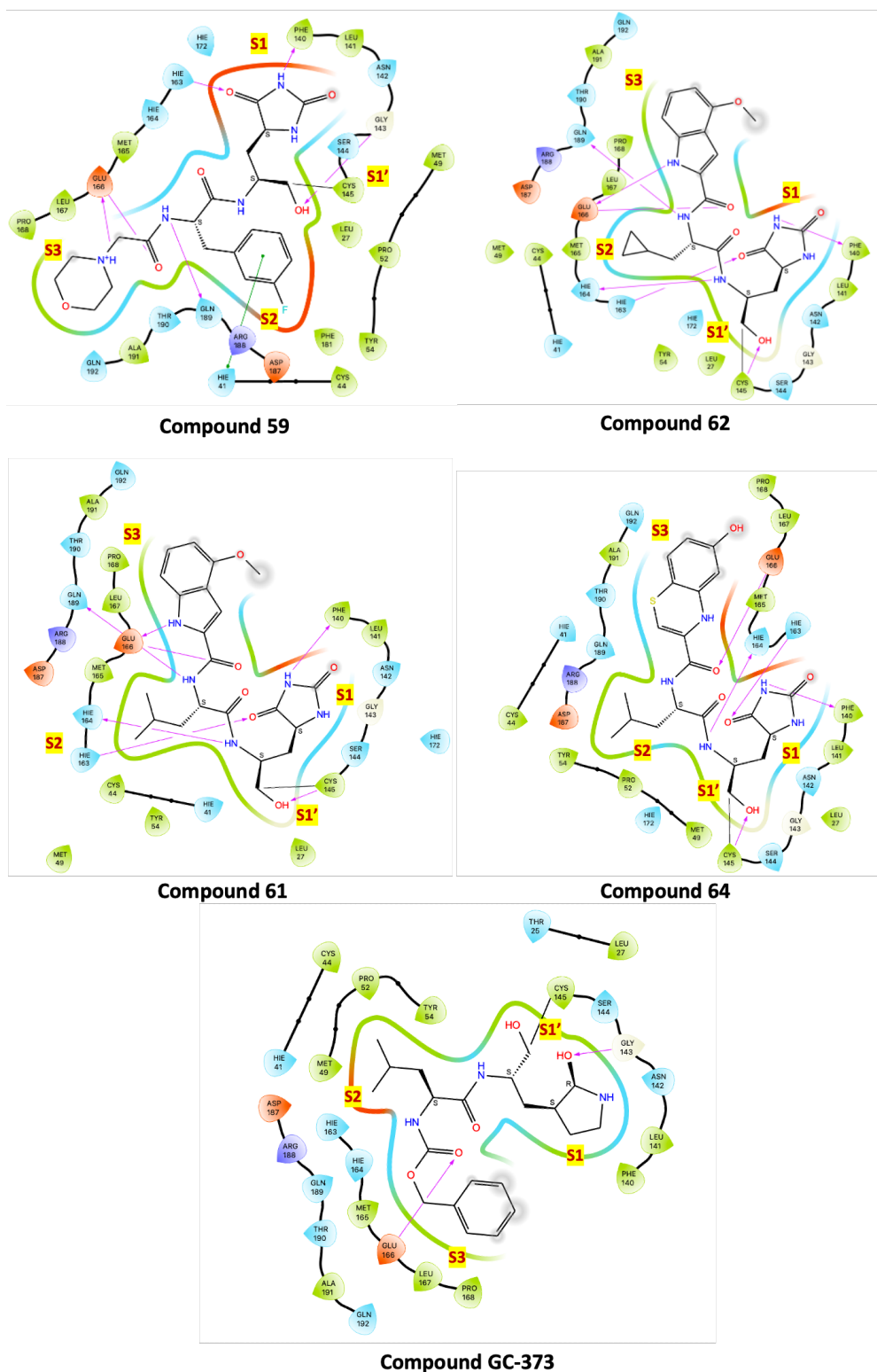
The analysis of 2D ligand interactions for derivatives **59**, **61**, **62**, and **64**, Figure 60, revealed important details about their binding properties and interaction patterns with the SARS-CoV-2 M<sup>PRO</sup>.

Once the covalent bond between the aldehydic warhead and the Cys<sup>145</sup> in the S1' pocket was established, the positioning of the P1, P2, and P3 moieties was examined.

The hydantoin moiety, which is recurrent in P1 of the top-ranking compounds, is capable of stabilizing hydrogen bonds with several amino acids in the S1 pocket, including Phe<sup>140</sup>, His<sup>163</sup>, His<sup>164</sup>, Glu<sup>166</sup>, His<sup>172</sup>. Specifically, two crucial interactions were noted: the carboxylic backbone group of Phe<sup>140</sup> forms a hydrogen bond with the NH group of the hydantoin moiety (Phe<sup>140</sup>-C=O---H-N-hydantoin), and the imidazole NH group of His<sup>163</sup> interacts with the Oxygen atom of the hydantoin moiety (His<sup>163</sup>-N-H---O=C-hydantoin). These interactions suggest that the hydantoin moiety could mimic the Gln residue of natural substrates, providing a potential alternative to the extensively studied  $\gamma$ -lactam ring.

In the S2 pocket, which is mainly composed of hydrophobic amino acids such as Met<sup>49</sup>, Tyr<sup>54</sup>, Met<sup>165</sup>, Leu<sup>167</sup>, Pro<sup>168</sup>, Val<sup>186</sup>, Asp<sup>187</sup>, Arg<sup>188</sup>, there is considerable flexibility allowing for the binding of both small and bulky aromatic/alkyl groups. In compounds **59**, **61**, **62**, and **64**, cyclopropyl, isoleucine, and phenyl moieties are well-stabilized in the S2 subregion through hydrophobic interactions.

For interactions with the S3 subregion, compounds **61**, **62**, and **64** feature aromatic moieties in P3, while **59** a morpholine group, that interact with amino acids like Gln<sup>189</sup>, Leu<sup>167</sup>, Pro<sup>168</sup>, Gly<sup>251</sup>, and Asp<sup>187</sup>. Overall, the identified compounds exhibit about 5 to 7 interactions with key residues in the binding pockets, in contrast to the reference compound **GC-373**, which shows only 2 interactions involving Glu<sup>166</sup> and Gly<sup>143</sup>. This suggests that the new derivatives have a more extensive interaction profile, which may contribute to their enhanced binding efficacy.

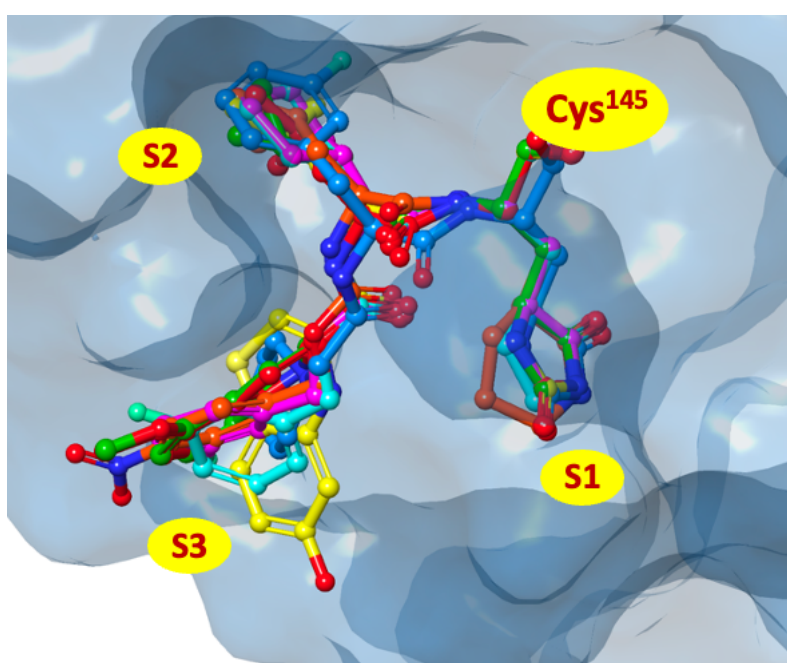


**Figure 60.** 2D ligand interaction maps for compounds 64, 66, 67, 69, and the reference GC-373.

The same trend is observed for compounds 57, 58, and 65, as shown in the 2D ligand interaction maps reported in Figure S1 of the Supplementary Material, while compounds 63,

**65, 67-69** demonstrated a lower number of interactions within the SARS-CoV-2 M<sup>PRO</sup> binding site.

Moreover, the analysis of the 3D binding poses for derivatives **57-62**, and **64**, as illustrated in Figure 61, reveals a significant overlap among the poses. This overlap indicates a redundancy in the positioning of the key elements of these small molecules within the four sub-pockets of the binding site. The consistent positioning of these elements results in the formation of multiple interactions across the sub-pockets, suggesting a potentially similar binding mechanism or interaction pattern among these derivatives.



**Figure 61.** 3D overlaps of compounds **57-62**, and **64** at the SARS-CoV-2 M<sup>PRO</sup> binding site.

Through a powerful blend of ligand- and structure-based approaches, supported by rigorous statistical analysis, this protocol has successfully identified new covalent peptidomimetic inhibitors with aldehydic warheads, demonstrating significant potential as SARS-CoV-2 M<sup>PRO</sup> inhibitors.



#### 4. Conclusions and Future Perspectives: Computational Approaches in SARS-CoV-2 M<sup>PRO</sup> Inhibitors Discovery

*In silico* simulations of complex systems are becoming increasingly popular in drug development and clinical research. Computational models and simulations offer significant advantages over traditional human-based clinical trials in both operational efficacy and therapeutic outcomes. They provide tools for qualitatively and quantitatively evaluating various treatments for specific diseases, leading to more practical and economical experiments. In the search for effective COVID-19 treatments, computational approaches have become essential in discovering and developing SARS-CoV-2 M<sup>PRO</sup> inhibitors. These *in silico* techniques allow researchers to model and predict interactions between potential inhibitors and the M<sup>PRO</sup> enzyme, facilitating the identification of promising drug candidates and optimizing their efficacy and selectivity. By employing various computational methodologies, scientists can accelerate the drug discovery process, reduce experimental costs, and enhance the precision of target identification and lead optimization.

Based on these premises, this PhD Thesis is aimed at developing and proposing new *in silico* protocols for the identification of diverse anti-SARS-CoV-2 agents targeted to M<sup>PRO</sup>. To this end, three different inhibition strategies were explored: non-covalent inhibition, covalent inhibition through non peptidomimetic inhibitors, and covalent inhibition through peptidomimetic inhibitors.

- Non-covalent inhibition strategy: this strategy focused on the design and identification of inhibitors capable of binding the catalytic site and/or allosteric sites of M<sup>PRO</sup>. In my research project, we identified a set of ethyl 3-benzoylamino-5-[(1H-imidazol-4-yl-methyl)-amino]-benzo[b]thiophene-2-carboxylates **2** and ethyl 3-benzoylamino-5-[(1H-imidazol-4-yl-methyl)-amino]-benzo[b]furan-2-carboxylates **3** as potential SARS-CoV-2 M<sup>PRO</sup> inhibitors, through a hierarchical and hybrid virtual screening. In details, the BPT, available in DRUDIT allows to filter a large in-house structure database, identifying the set of small molecules with high affinity against the SARS-CoV-2 M<sup>PRO</sup> catalytic binding site. ADME properties of the selected compounds were investigated through the SwissADME tool and IFD studies were performed on the catalytic site to confirm DRUDIT prediction. Moreover, aiming at evaluating the possibility of a dual binding mechanism of action, the identified hits were further investigated by means

of PCA, and IFD into the dimerization site. Compounds **2i-l** also exhibited promising  $IC_{50}$  when *in vitro* evaluated as inhibitors of the catalytic site, among them **2i** ( $IC_{50}$  values of 70.4) and **2l** ( $IC_{50}$  values of 45.9) are promising lead compounds for further development as antiviral agents with a dual binding activity.

- Covalent inhibition strategy through non-peptidomimetic inhibitors: in this approach, we focused on the rational design and synthesis of compound with non peptidomimetic scaffold, capable of covalently blocking the catalytic cysteine residue (Cys<sup>145</sup>) of M<sup>PRO</sup>, thus permanently inactivating the enzyme. In details we synthesized compounds **45a-m** and identified derivative **45g** as the most promising of the series, with  $IC_{50}$  value of  $30 \pm 6.6 \mu\text{M}$ . These esters offer the advantage of long-lasting inhibition, becoming a potent strategy for disrupting viral replication.

Future analysis, such as covalent adducts formation and Lead Optimization process, will be performed, especially starting from compound **45g**.

- Covalent inhibition strategy through peptidomimetic inhibitors: peptidomimetic inhibitors are developed to closely mimic the structure of the natural viral polyproteins that are substrates of M<sup>PRO</sup>, ensuring highly specific interactions. Utilizing the advanced CombiGlide Tool within the Maestro suite, a highly focused combinatorial library of 450 rationally designed peptidomimetic compounds with aldehydic warheads was generated. IFD studies targeted at the catalytic site of M<sup>PRO</sup> enabled the first refinement, narrowing the field to 388 compounds for subsequent multivariate statistical analysis. Of these, 14 standout compounds, identified through PCA, were meticulously evaluated using CovDock studies to assess their covalent binding capabilities. The results, especially for compounds **57-62**, and **64**, reveal significantly higher affinity compared to known inhibitors, affirming the validity of the adopted design strategy. The crucial interactions between the P1, P2, and P3 moieties of the selected compounds and the M<sup>PRO</sup> active site indicate effective molecular mimicry of natural substrates, suggesting the potential for developing highly selective and potent inhibitors.

Our findings demonstrate that the combination of *in silico* methods and advanced computational tools represents a strategic and highly promising path for the identification of new antiviral drugs. My thesis took the advantages of our in-house ligand based BPT which allowed us to screen an enormous ligands library in negligible computational time and with

no need for particularly high-performance hardware. This tool integrated with both structure-based techniques and, interestingly, multivariate statistical analysis, has been applied to evaluate potential new SARS-CoV-2 M<sup>PRO</sup> inhibitors. This research project provides a foundation for future research and design of selective antiviral agents to combat COVID-19.

## 5. Materials and Methods

### 5.1. *In silico* studies

#### 5.1.1. Ligand-based studies

The DRUDIT web-service operates through four servers, each of which can perform more than ten jobs simultaneously, and several software modules implemented in C and JAVA running on MacOS Mojave.

##### 5.1.1.1. MOLDESTO: a new software for molecular descriptors calculation

MOLDESTO, as we described in a previous work [108], is a software tool implemented in DRUDIT that represents the evolution of the expertise of the research group in the calculation/ manipulation of molecular descriptors [158]. It is currently able to calculate more than 1000 1D, 2D, and 3D molecular descriptors for each input structure. The input structures can be drawn directly in the web interface or uploaded as commonly used molecule file formats as external files (e.g. SMILES, SDF, Inchi, Mdl, Mol2). The software is provided with a caching system to boost the calculation speed of previously submitted structures.

##### 5.1.1.2. Biotarget Predictor Tool

The Biotarget Finder Module provides prediction of the binding affinity between candidate molecules and the specified biological target.

The tool was used to screen small molecule drug candidates as SARS-CoV-2 M<sup>PRO</sup> inhibitors [108]. The template of SARS-CoV-2 M<sup>PRO</sup> was created using a set of well-known drugs to perform molecular docking studies at catalytic site of SARS-CoV-2 M<sup>PRO</sup>. It was uploaded in DRUDIT, and the default DRUDIT parameters (N=500, Z=50, G=a) were used [9,108]. In accordance with the first phase of the *in silico* workflow in section 3.1.1.1, the in-house database was uploaded in DRUDIT and submitted to the BPT. The output results were obtained as DAS for each structure, reflecting, the binding affinity of compounds against the SARS-CoV-2 M<sup>PRO</sup> catalytic site.

### 5.1.2. Combiglide: Design of a combinatorial library

The combinatorial ligand library, section 3.2.3.1., was generated using CombiGlide to explore potential chemical space efficiently. After the design of a retrosynthetic scheme, the following steps were involved in the library construction:

#### 5.1.2.1. Reagent Preparation Panel

In this panel it is possible to prepare reagent files for use in combinatorial library enumeration. Reagent preparation ensures that the input structures are all-atom, 3D structures, and that they have the appropriate information stored with them to construct the molecules that are used in screening or are added to the library. Additionally, the main task in this panel is to select a reagent type (a functional group) and to identify the bond in the functional group that is replaced when the reagent is added to the core. A reagent preparation job for each reagent type was run.

#### 5.1.2.2. Combinatorial Library Enumeration Panel

In this tab, the core-containing molecule was set up by importing a core-containing molecule and defining its attachment bonds.

### 5.1.3. Structure-based studies

The preparation process of ligands and protein-ligand complexes used for *in silico* studies has been performed as detailed below:

#### 5.1.3.1. Ligand Preparation

The ligands for docking were prepared through the LigPrep tool, available in the Maestro Suite, Schrödinger software [204]. For each ligand, all possible tautomers and stereoisomers were generated for a pH of  $7.0 \pm 0.4$ , using default setting, through the Epik ionization method [205]. Consequently, the integrated Optimized Potentials for Liquid Simulations (OPLS) 2005 force field was used to minimize the energy status of the ligands [206].

### 5.1.3.2. Protein Preparation

The crystal structures of SARS-CoV-2 M<sup>PRO</sup> (PDB codes 7VH8 [116], 7AXM [83], and 6WTK [126]) were downloaded from the Protein Data Bank [207,208]. As regard PDB codes 7VH8 and 6WTK a first breakup of the covalent bound between the cocrystal ligands and the Cys<sup>145</sup> was carried out. Successively, the proteins were prepared using the Protein Preparation Wizard, in the Schrödinger software, with the default setting [209]. In detail, bond orders were assigned, including Het group, hydrogen atoms were added, all water molecules were deleted, and protonation of the heteroatom states were carried out using the Epik-tool (with the pH set at biologically relevant values, i.e., at  $7.0 \pm 0.4$ ). The H-bond network was then optimized. The structure was finally subjected to a restrained energy minimization step (RMSD of the atom displacement for terminating the minimization was 0.3 Å), using the OPLS 2005 force field [206].

### 5.1.3.3. Docking Validation

Molecular Docking studies were executed and scored by using the Glide module, available in the Schrödinger Suite program package. The receptor grids were obtained through assignment the original ligands (*nirmatrelvir*, *pelitinib*, and GC-373 for PDB codes 7VH8 [116], 7AXM [83], and 6WTK [126], respectively) as the centroid of the grid boxes. Extra Precision (XP) mode, as scoring function, was used to dock the generated 3D conformers into the receptor model. The post-docking minimization step was performed with a total of 5 poses for each ligand conformer, and a maximum of 2 docking poses were generated per ligand conformer. The proposed docking procedure was able to re-dock the original ligands within the receptor-binding pockets with  $\text{RMSD} < 0.51 \text{ \AA}$ . Table 17 shows all the parameters combinations used for RMSD value optimization for the proteins. In detail, Radii Van der Waals Scaling allows to temporarily remove active-site residue side chains. By default, the scaling factor is 0.50 for the receptor and 0.50 for the ligand, with a partial charge threshold of 0.15. Removing side chains from active site residues provides more room for ligand docking, so the receptor does not need to be quite as soft. The side chains are restored after docking. The Side chain optimization make possible to reduce the distance from the ligand that defines residues for refinement. In general, the optimal value for this parameter is default set 5.0 Å, ensuring that optimize side chains is selected.

The Energy Minimization parameters controls the minimization protocol through the distance-dependent dielectric constant (the optimum is to set the protein/ligand dielectric constants to values of 1-2, default setting at 2) and the maximum number of minimization steps (by default fixed to 100) [206].

**Table 17.** Parameters Docking Combinations for RMDS value optimization.

Radii Van der Waals Scaling			Side chain optimization	Energy Minimization		RMSD		
<i>Receptor Van der Waals Scaling</i>	<i>Ligand Van der Waals Scaling</i>	<i>Partial charge cut-off</i>	<i>Residue Refinement</i>	<i>Distance - dependent dielectric constant</i>	<i>Maximum number of minimization steps</i>	<i>7VH8</i>	<i>7AXM</i>	<i>6WTK</i>
1.50	1.50	0.75	3 Å	0.5	20	0.87 Å	0.86 Å	0.88 Å
1.25	1.25	0.50	3.5 Å	0.75	40	0.73 Å	0.75 Å	0.75 Å
1.00	1.00	0.35	4 Å	1.00	60	0.66 Å	0.68 Å	0.66 Å
0.75	0.75	0.25	4.5 Å	1.50	80	0.59 Å	0.57 Å	0.57 Å
0.50	0.50	0.15	5 Å	2.0	100	0.51 Å	0.51 Å	0.51 Å

#### 5.1.3.4. Induced Fit Docking

Induced Fit Docking simulation was performed using the IFD application, an accurate and robust Schrödinger technology that accounts for both ligand and receptor flexibility [210,211]. Schrödinger Induced Fit Docking validated protocol was applied by using SARS-CoV-2 M<sup>PRO</sup> protein from the PDB (PDB codes 7VH8 [116] and 7AXM [83]), previously refined by the Protein Preparation module. The IFD score (IFD score = 1.0 Glide Gscore + 0.05 Prime Energy), which includes protein–ligand interaction energy and system total energy, was calculated and used to rank the IFD poses. The more positive in modulus was the IFD score, the more favorable was the binding.

#### 5.1.3.5. Covalent Docking

Covalent Docking was carried out to dock a set of peptidomimetic ligands within the SARS-CoV-2 M<sup>PRO</sup> binding site (PDB code 6WTK [126]). Unlike classical docking, where interactions are mainly non-covalent, covalent docking requires specific modeling of the chemical reaction process occurring between a functional group of the ligand and a reactive residue on the protein. The covalent docking process was performed using Maestro Schrödinger, which

incorporates an algorithm capable of handling both the pre-covalent phase (geometric alignment and orientation of the ligand in the active site) and the post-covalent phase (simulating the formation of the chemical bond).

The workflow adopted includes the following main steps:

- Identification of the reactive residue: the binding site and the reactive residue (Cys<sup>145</sup>) were identified in the SARS-CoV-2 M<sup>PRO</sup> binding site (PDB code 6WTK [126]).
- Reaction Type selection: the reaction type for the formation of covalent bond between ligand and receptor was selected. The reaction type defines the functional group on the ligand and the operations that must be performed on the structures to create the covalent bond (Nucleophilic Addition to a Double Bond).
- Pre-covalent docking: a preliminary docking was performed to optimally position the ligand in the binding site before covalent bond formation. The interaction energy was evaluated using an algorithm based on a scoring function that considers non-covalent interactions.
- Covalent bond formation: a chemical model was applied to simulate the formation of the covalent bond between the nucleophilic group of the reactive residue and the ligand.
- Evaluation of the protein-ligand complex: the stability of the covalent protein-ligand complex was assessed through a scoring function specific to covalent docking, accounting for both covalent bond energy and residual non-covalent interactions. Additionally, a short molecular dynamics simulation was performed to further validate and optimize the geometry of the final complex.

## 5.2. Multivariate Statistical Analysis

PCA, one of the most-widely used multivariate exploratory techniques, enables a drastic dimensionality reduction of an original raw data, transforming the original matrix to a new one, whose set of variables, termed as Principal Components (PCs), appear to be ordered with descending importance in terms of variance. Principal Components Analysis can be highly useful for data classification and pattern recognition. In this work DRUDIT was used to obtain the original matrix of objects versus variables (Supplementary Material S1, Matrices S1, S3,



and S4), and free version TIBCO Statistica® software was used to perform Principal Component Analysis.

As regard section 3.1.1.4. Grubb's test, also known as ESD method (Extreme Studentized Seviat), was performed. In details, the outliers were determined by evaluating singularly those compounds outside the red circle (Figure 16) in comparison with the cluster of molecules closest to *pelitinib*. The identified outliers were not included in the step after.

## 5.3. Chemistry

### 5.3.1. General Information

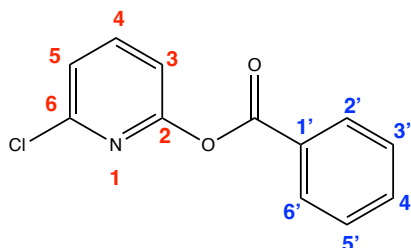
Unless otherwise indicated, all reagents and solvents were purchased from commercial sources and used without further purification. All melting points (°C) were determined on a Büchi Tottoli capillary apparatus and are uncorrected; IR spectra were determined in bromoform with a Jasco FT/IR 5300 spectrophotometer. <sup>1</sup>H NMR, <sup>13</sup>C NMR spectra were recorded, at 400 and 100 MHz respectively, in CDCl<sub>3</sub> or DMSO-d<sub>6</sub> solution, using a Bruker AC-E series 200 MHz spectrometer. Chemical shifts values are given in ppm and referred as the internal standard to tetramethylsilane (TMS). The following abbreviations are used: br s = broad signal, s = singlet, d = doublet, t = triplet, q = quartet, m = multiplet, rt = room temperature. The NMR of compounds are reported in the Supplementary Figures S2-S27. The purity of all compounds screened in biological assays was determined to be >95% by HPLC/MS analysis. Mass spectra were performed using a GC and MS Shimadzu QP5050 with EI (75 ev). Microanalyses agreed with theoretical values ±0.4%. Thin layer chromatography was performed on precoated (0.25 mm) silica gel GF254 plates, compounds were detected with 254 nm UV lamp. Column chromatography was performed with Merck silica gel ASTM (230 and 400 mesh), or with a Biotage FLASH40i chromatography module (prepacked cartridge system). 2-Chloro-6-Hydroxy-pyridine (**46**), benzoyl chlorides **47** and benzoic acids **48** are commercially available. Final products were obtained by two different synthesis routes.

### 5.3.2. General procedure for the synthesis of 6-chloropyridin-2-yl benzoate (45a-m)

*Method A:* To a stirring solution of 2-chloro-6-hydroxy-pyridine **46** (1 eq.) in pyridine (1.2 eq.) cooled at 0°C, a solution of the appropriate benzoyl chloride in DCM (**47**, 1.2 eq.) was added.

The reaction mixture was allowed to stir at RT (25 °C) for 12-24 h. Then the solvent was evaporated under vacuum and the obtained crude was poured onto stirred water/ice, extracted with DCM and washed with a solution of 1M HCl. The organic phase was dried over Na<sub>2</sub>SO<sub>4</sub> and evaporated under vacuum. The obtained crude was purified through silica gel chromatography using a mixture 10:1 petroleum ether/EtOAc in gradient as eluant to afford the title compound as white powder.

*Method B:* To a stirring solution of 2-chloro-6-hydroxy-pyridine **46** (1 eq.) in anhydrous DCM, the corresponding carboxylic acid (**48** 1 eq.) and DCC (1.5 eq.) were added under inert atmosphere (N<sub>2</sub>). To the resulting mixture cooled at 0°C, a catalytic amount of 4-dimethylaminopyridine (DMAP, 0.5 eq.) was carefully added, and the reaction was then allowed to stir at RT for 12-24 h. The solvent was evaporated under vacuum and the obtained crude was poured onto stirred water/ice, extracted with DCM. The organic phase was dried over Na<sub>2</sub>SO<sub>4</sub> and evaporated under vacuum. The obtained crude was purified through silica gel chromatography using a mixture 10:1 petroleum ether/EtOAc in gradient as eluant to afford the title compound as white powder.



### 5.3.2.1. Synthesis of 6-chloropyridin-2-yl 5-chloro-2-nitrobenzoate (**45a**)

Yield 85-95%. Mp 60-65 °C. <sup>1</sup>H NMR (DMSO-*d*<sub>6</sub>) δ: 7.46 (dd, 1H, J = 8.0, 0.6 Hz, H-3), 7.63 (dd, 1H, J = 7.8, 0.6 Hz, H-5), 8.03 (dd, 1H, J = 8.8, 2.3 Hz, H-4'), 8.14 (t, 1H, J = 7.9 Hz, H-4), 8.26 (d, 1H, J = 8.8 Hz, H-3'), 8.29 (d, 1H, J = 2.3 Hz, H-6). <sup>13</sup>C NMR (DMSO-*d*<sub>6</sub>) δ: 115.9, 124.3, 127.1, 127.3, 130.6, 133.9, 139.4, 144.2, 146.3, 148.7, 156.2, 162.0. HRMS-ESI [(M+H)<sup>+</sup>]: m/z calculated for C<sub>12</sub>H<sub>6</sub>Cl<sub>2</sub>N<sub>2</sub>O<sub>4</sub>: 312.9777; found: 312.9778.

### 5.3.2.2. Synthesis of 6-chloropyridin-2-yl 2-amino-3,5-dichlorobenzoate (**45b**)

Yield 85-95%. <sup>1</sup>H NMR (DMSO-*d*<sub>6</sub>) δ: 6.95 (br s, 2H, NH<sub>2</sub>), 7.43 (dd, 1H, J = 8.0, 0.7 Hz, H-3), 7.58 (dd, 1H, J = 7.8, 0.6 Hz, H-5), 7.80 (d, 1H, J = 2.5 Hz, H-4'), 7.93 (d, 1H, J = 2.5 Hz, H-6'),

8.09 (t, 1H, J = 7.9 Hz, H-4). <sup>13</sup>C NMR (DMSO-*d*<sub>6</sub>) δ: 109.7, 116.6 (CH), 118.5, 121.0, 123.6 (CH), 125.9, 129.7 (CH), 135.0 (CH), 143.8 (CH), 147.3, 148.5, 164.4. HRMS-ESI [(M+H)<sup>+</sup>]: m/z calculated for C<sub>12</sub>H<sub>7</sub>Cl<sub>3</sub>N<sub>2</sub>O<sub>2</sub>: 316.9646; found: 316.9647.

### 5.3.2.3. Synthesis of 6-chloropyridin-2-yl 3,4,5-trimethoxybenzoate (45c)

Yield 85-95%. Mp 140-145°C. <sup>1</sup>H NMR (DMSO-*d*<sub>6</sub>) δ: 3.79 (s, 3H, OCH<sub>3</sub>), 3.87 (s, 6H, 2 x OCH<sub>3</sub>), 7.38 – 7.44 (m, 3H, H-3, H-2', H-6'), 7.57 (d, 1H, J = 7.7 Hz, H-5), 8.08 (t, 1H, J = 7.9 Hz, H-4). <sup>13</sup>C NMR (DMSO-*d*<sub>6</sub>) δ: 56.7 (CH<sub>3</sub>), 60.8 (CH<sub>3</sub>), 107.8 (CH), 116.5 (CH), 123.3, 123.5 (CH), 143.3, 143.8 (CH), 148.6, 153.4, 157.4, 164.1. HRMS-ESI [(M+H)<sup>+</sup>]: m/z calculated for C<sub>15</sub>H<sub>14</sub>ClNO<sub>5</sub>: 324.0633; found: 324.0635.

### 5.3.2.4. Synthesis of 6-chloropyridin-2-yl 4-methoxybenzoate (45d)

Yield 85-95%. Mp 88°C. <sup>1</sup>H NMR (DMSO-*d*<sub>6</sub>) δ: 3.88 (s, 3H, OCH<sub>3</sub>), 7.14 (d, 2H, J = 9.0 Hz, H-3', H-5'), 7.40 (dd, 1H, J = 8.0, 0.6 Hz, H-3), 7.55 (dd, 1H, J = 7.8, 0.6 Hz, H-5), 8.02 – 8.13 (m, 3H, H-4, H-2', H-6'). <sup>13</sup>C NMR (DMSO-*d*<sub>6</sub>) δ: 56.2 (CH<sub>3</sub>), 114.9 (CH), 116.5 (CH), 120.5, 123.3 (CH), 132.7 (CH), 143.7 (CH), 148.5, 157.6, 164.1, 164.7. HRMS-ESI [(M+H)<sup>+</sup>]: m/z calculated for C<sub>13</sub>H<sub>10</sub>ClNO<sub>3</sub>: 264.0422; found: 264.0420.

### 5.3.2.5. Synthesis of 6-chloropyridin-2-yl 4-(trifluoromethyl)benzoate (45e)

Yield 85-95%. Mp 58-60 °C. <sup>1</sup>H NMR (DMSO-*d*<sub>6</sub>) δ: 7.49 (dd, 1H, J = 7.9, 0.6 Hz, H-3), 7.60 (dd, 1H, J = 7.8, 0.6 Hz, H-5), 8.01 (d, 2H, J = 8.3 Hz, H-3', H-5'), 8.11 (t, 1H, J = 7.9 Hz, H-4), 8.33 (d, 2H, J = 8.1 Hz, H-2', H-6'). <sup>13</sup>C NMR (DMSO-*d*<sub>6</sub>) δ: 116.4 (CH), 123.8 (CH), 124.1, 126.6 (CH), 131.3 (CH), 132.4, 134.3, 143.9 (CH), 148.6, 157.1, 163.4. HRMS-ESI [(M+H)<sup>+</sup>]: m/z calculated for C<sub>13</sub>H<sub>7</sub>ClF<sub>3</sub>NO<sub>2</sub>: 302.0190; found: 302.0190.

### 5.3.2.6. Synthesis of 6-chloropyridin-2-yl 4-methylbenzoate (45f)

Yield 85-95%. Mp 45-50 °C. <sup>1</sup>H NMR (DMSO-*d*<sub>6</sub>) δ: 2.43 (s, 3H, CH<sub>3</sub>), 7.39-7.47 (m, 3H, H-3, H-3', H-5'), 7.56 (d, 1H, J = 7.8 Hz, H-5), 8.02 (d, 2H, J = 8.2 Hz, H-2', H-6'), 8.07 (t, 1H, J = 7.9 Hz, H-4). <sup>13</sup>C NMR (DMSO-*d*<sub>6</sub>) δ: 21.8 (CH<sub>3</sub>), 116.5 (CH), 123.4 (CH), 125.7, 130.2 (CH), 130.5 (CH),

143.8 (CH), 145.7, 148.6, 157.5, 164.4. HRMS-ESI [(M+H)<sup>+</sup>]: m/z calculated for C<sub>13</sub>H<sub>10</sub>ClNO<sub>2</sub>: 248.0473; found: 248.0475.

#### 5.3.2.7. Synthesis of 6-chloropyridin-2-yl benzoate (45g)

Yield 85-95%. Mp 48-50 °C. <sup>1</sup>H NMR (DMSO-*d*<sub>6</sub>) δ: 7.44 (dd, 1H, J = 8.0, 0.6 Hz, H-3), 7.58 (dd, 1H, J = 7.8, 0.6 Hz, H-5), 7.60 – 7.66 (m, 2H, H-3', H-5'), 7.75 – 7.82 (m, 1H, H-4'), 8.09 (t, 1H, J = 7.9 Hz, H-4), 8.12 – 8.16 (m, 2H, H-2', H-6'). <sup>13</sup>C NMR (DMSO-*d*<sub>6</sub>) δ: 116.5 (CH), 123.5 (CH), 128.5, 129.6 (CH), 130.4 (CH), 135.0 (CH), 143.8 (CH), 148.6, 157.4, 164.5. HRMS-ESI [(M+H)<sup>+</sup>]: m/z calculated for C<sub>12</sub>H<sub>8</sub>ClNO<sub>2</sub>: 234.0316; found: 234.0314.

#### 5.3.2.8. Synthesis of 6-chloropyridin-2-yl 4-(dimethylamino)benzoate (45h)

Yield 85-95%. Mp 105-120 °C. <sup>1</sup>H NMR (DMSO-*d*<sub>6</sub>) δ: 3.05 (s, 6H, 2 x CH<sub>3</sub>), 6.80 (d, 2H, J = 9.1 Hz, H-3', H-5'), 7.34 (dd, 1H, J = 8.0, 0.6 Hz, H-3), 7.51 (dd, 1H, J = 7.8, 0.6 Hz, H-5), 7.91 (d, 2H, J = 9.0 Hz, H-2', H-6'), 8.03 (t, 1H, J = 7.9 Hz, H-4). <sup>13</sup>C NMR (DMSO-*d*<sub>6</sub>) δ: 40.1 (CH<sub>3</sub>), 111.5 (CH), 113.8, 116.6 (CH), 122.8 (CH), 132.3 (CH), 143.4 (CH), 148.5, 154.5, 158.0, 164.4. HRMS-ESI [(M+H)<sup>+</sup>]: m/z calculated for C<sub>14</sub>H<sub>13</sub>ClN<sub>2</sub>O<sub>2</sub>: 277.0738; found: 277.0740.

#### 5.3.2.9. Synthesis of 6-chloropyridin-2-yl 4-fluorobenzoate (45i)

Yield 85-95%. Mp 67-72 °C. <sup>1</sup>H NMR (DMSO-*d*<sub>6</sub>) δ: 7.41 – 7.50 (m, 3H, H-3, H-3', H-5'), 7.58 (d, 1H, J = 7.8 Hz, H-5), 8.09 (t, 1H, J = 7.9 Hz, H-4), 8.17 – 8.25 (m, 2H, H-2', H-6'). <sup>13</sup>C NMR (DMSO-*d*<sub>6</sub>) δ: 116.5 (CH), 116.8 (d, J<sub>C-F</sub> = 22.3 Hz, CH), 123.5 (CH), 125.1, 133.5 (d, J<sub>C-F</sub> = 9.8 Hz, CH), 143.8 (CH), 148.6, 157.3, 163.5, 166.3 (d, J = 253.8 Hz). HRMS-ESI [(M+H)<sup>+</sup>]: m/z calculated for C<sub>12</sub>H<sub>7</sub>ClFNO<sub>2</sub>: 252.0222; found: 252.0220.

#### 5.3.2.10. Synthesis of 6-chloropyridin-2-yl 4-bromobenzoate (45j)

Yield 85-95%. Mp 129-133 °C. <sup>1</sup>H NMR (DMSO-*d*<sub>6</sub>) δ: 7.45 (dd, 1H, J = 8.0, 0.6 Hz, H-3), 7.58 (dd, 1H, J = 7.8, 0.6 Hz, H-5), 7.84 (d, 2H, J = 8.6 Hz, H-3', H-5'), 8.05 (d, 2H, J = 8.6 Hz, H-2', H-6'), 8.09 (t, 1H, J = 7.9 Hz, H-4). <sup>13</sup>C NMR (DMSO-*d*<sub>6</sub>) δ: 116.5 (CH), 123.6 (CH), 127.8, 129.3,

132.3 (CH), 132.8 (CH), 143.9 (CH), 148.6, 157.2, 163.8. HRMS-ESI [(M+H)<sup>+</sup>]: m/z calculated for C<sub>12</sub>H<sub>7</sub>BrClNO<sub>2</sub>: 311.9421; found: 311.9420.

#### 5.3.2.11. Synthesis of 6-chloropyridin-2-yl 2,3-dichlorobenzoate (45k)

Yield 85-95%. Mp 98-100 °C. <sup>1</sup>H NMR (DMSO-*d*<sub>6</sub>) δ: 7.49 (dd, 1H, J = 8.0, 0.6 Hz, H-3), 7.55 – 7.64 (m, 2H, H-5, H-5'), 7.98 (dd, 1H, J = 8.1, 1.5 Hz, H-6'), 8.08 (dd, 1H, J = 7.8, 1.6 Hz, H-4'), 8.12 (t, 1H, J = 7.9 Hz, H-4). <sup>13</sup>C NMR (DMSO-*d*<sub>6</sub>) δ: 116.3 (CH), 123.9 (CH), 129.2 (CH), 130.8, 130.9 (CH), 131.2, 134.1, 135.2 (CH), 144.0 (CH), 148.7, 156.8, 162.6. HRMS-ESI [(M+H)<sup>+</sup>]: m/z calculated for C<sub>12</sub>H<sub>6</sub>Cl<sub>3</sub>NO<sub>2</sub>: 301.9537; found: 301.9535.

#### 5.3.2.12. Synthesis of 6-chloropyridin-2-yl 4-chlorobenzoate (45l)

Yield 85-95%. Mp 115-120 °C. <sup>1</sup>H NMR (DMSO-*d*<sub>6</sub>) δ: 7.45 (dd, 1H, J = 8.0, 0.7 Hz, H-3), 7.58 (dd, 1H, J = 7.8, 0.6 Hz, H-5), 7.70 (d, 2H, J = 8.7 Hz, H-3', H-5'), 8.09 (t, 1H, J = 7.9 Hz, H-4), 8.13 (d, 2H, J = 8.7 Hz, H-2', H-6'). <sup>13</sup>C NMR (DMSO-*d*<sub>6</sub>) δ: 116.5 (CH), 123.6 (CH), 127.4, 129.8 (CH), 132.3 (CH), 140.1, 143.9 (CH), 148.6, 157.2, 163.6. HRMS-ESI [(M+H)<sup>+</sup>]: m/z calculated for C<sub>12</sub>H<sub>7</sub>Cl<sub>2</sub>NO<sub>2</sub>: 267.9927; found: 267.9927.

#### 5.3.2.13. Synthesis of 6-chloropyridin-2-yl 3-chloro-4-fluorobenzoate (45m)

Yield 85-95%. <sup>1</sup>H NMR (DMSO-*d*<sub>6</sub>) δ: 7.45 (dd, 1H, J = 8.0, 0.6 Hz, H-3), 7.59 (dd, 1H, J = 7.8, 0.6 Hz, H-5), 7.68 (t, 1H, J = 8.9 Hz, H-5'), 8.10 (t, 1H, J = 7.9 Hz, H-4), 8.16 (ddd, 1H, J = 8.7, 4.7, 2.2 Hz, H-6'), 8.29 (dd, 1H, J = 7.1, 2.2 Hz, H-2'). <sup>13</sup>C NMR (DMSO-*d*<sub>6</sub>) δ: 116.4 (CH), 118.3 (d, J<sub>C-F</sub> = 22.1 Hz, CH), 121.2 (d, J<sub>C-F</sub> = 18.4 Hz), 123.7 (CH), 126.4 (d, J<sub>C-F</sub> = 3.3 Hz), 131.9 (d, J<sub>C-F</sub> = 9.0 Hz, CH), 132.9 (CH), 143.9 (CH), 148.6, 157.0, 161.4 (d, J<sub>C-F</sub> = 256 Hz), 162.5. HRMS-ESI [(M+H)<sup>+</sup>]: m/z calculated for C<sub>12</sub>H<sub>6</sub>Cl<sub>2</sub>FNO<sub>2</sub>: 285.9832; found: 285.9830.

### 5.4. Biological Procedures

#### 5.4.1. Antiviral Assay

The M<sup>PRO</sup> activity was assayed fluorometrically with the kit SensoLyte 520 SARS-CoV-2 3CL Protease Activity Assay Kit (Anaspec, DBA Italy), essentially as described by the manufacturer.

Briefly, serially diluted drugs (10 $\mu$ L), ranging from 5000 $\mu$ M to 1 $\mu$ M, were added in a black walled flat microplate to the 3CL enzyme provided (40 $\mu$ L), then 50 $\mu$ L of the substrate (HiLyte™ Fluor 488/QXL® 520) was added and the plate incubated at 37°C for 1 hour. Then fluorescence was measured at Ex490nm and Em520 on a CLARIOStar multimode plate reader (BMG Labtech, Milan, Italy). Control wells with no inhibitors have been set, background fluorescence (substrate only) have been subtracted from all the wells, and a positive control (GC376 10nM) has been set. Results have been expressed as percentage of control wells set as 100% with formula  $(F_{\text{sample}}/F_{\text{control}})/100$ . IC<sub>50</sub> were calculated using Prism 6 following the Hill equation (GraphPad software, CA, USA).

6. **Appendix: Visiting research period at the University of Buckingham,  
Oxford Drug Design (October 2023 – May 2024), Supervisor: Prof. Paul  
W. Finn**

## **6.1. Machine learning model for predicting molecular activity using Molecular Descriptors and ElectroShape Descriptors**

In the final section of this PhD thesis, a dedicated chapter presents the results of the research conducted during a visiting research period at the University of Buckingham, supervised by Prof. Paul W. Finn, CEO at Oxford Drug Design (Oxford, October 2023 - April 2024). The focus of this project was the development of a Machine Learning (ML) model aimed at discriminating between active and inactive compounds for specific biological targets.

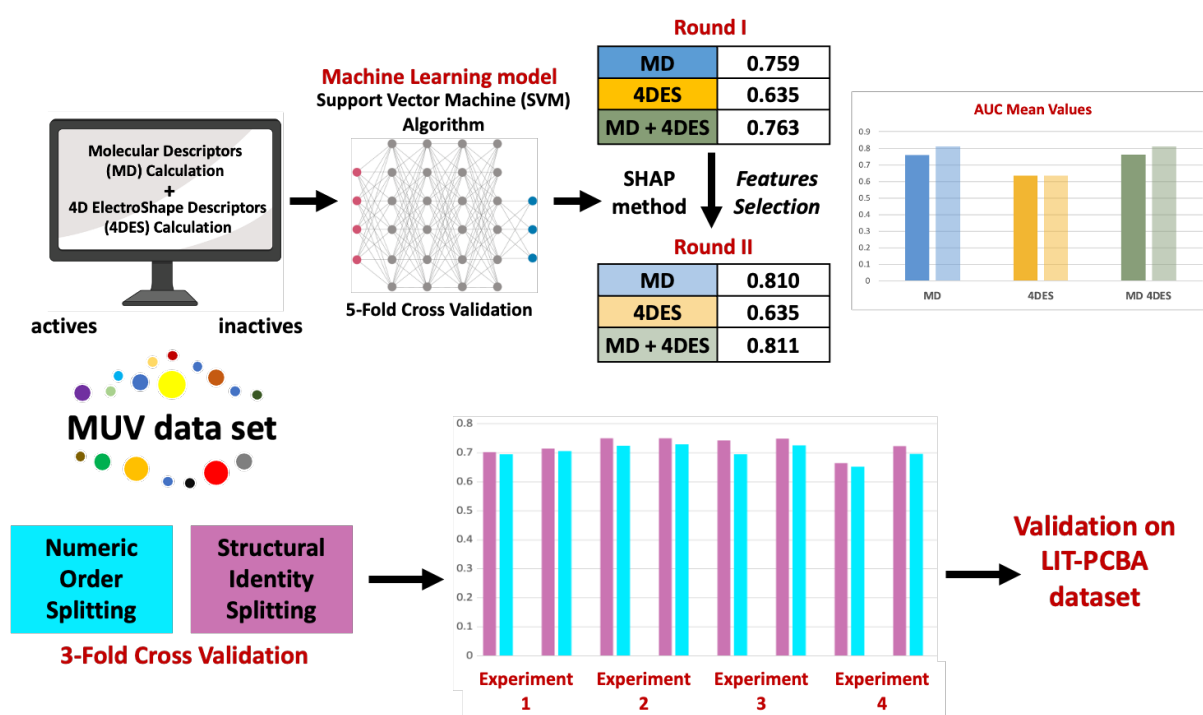
ML algorithms have been employed to analyze large datasets of molecular structures and bioactivity data, enabling the rapid identification of potential inhibitors for various proteins [212-219]. By training these models on known inhibitors, researchers can predict the binding affinity and inhibitory potential of novel compounds with high accuracy. These ML-driven approaches not only accelerate the drug discovery process but also enable the exploration of chemical space beyond the capability of traditional High-Throughput Screening (HTS) methods.

Our approach utilized Molecular Descriptors (MD), computed via MOLDESTO [108], alongside 4-Dimensional ElectroShape Descriptors (4DES) [220] to represent small molecules in the Maximum Unbiased Validation (MUV) dataset [221]. MD capture various physicochemical properties of small molecules, while 4DES are more advanced descriptors that incorporate three-dimensional spatial and electronic properties, offering a richer representation for detecting subtle molecular differences. The MUV dataset, which includes assay data for 17 targets, is designed to minimize analogue bias and artificial enrichment.

We developed a Support Vector Machine (SVM) model and applied 5-fold cross-validation to assess its performance using metrics such as Area Under the Curve (AUC) and the 1% Enrichment Factor (1%EF). Initial results showed that models using only MD achieved an AUC of  $0.759 \pm 0.099$ , while those using 4DES alone achieved  $0.635 \pm 0.113$ . Combining MD and 4DES slightly improved performance, reaching an AUC of  $0.763 \pm 0.099$ . Further refinement, employing SHapley Additive exPlanations (SHAP) for feature selection, enhanced the model's performance, increasing the AUC to  $0.810 \pm 0.082$  for MD alone and  $0.811 \pm 0.084$  for the combined descriptors. These findings underscore the importance of feature selection in improving model accuracy. Additionally, we conducted a 3-fold cross-validation by splitting the MUV dataset into three groups based on target numerical codes and FASTA sequence



identity. This demonstrated the model's robustness to discriminate between actives and inactives across different test scenarios. Finally, we evaluated the model on the LIT-PCBA dataset [222], achieving promising performance. These findings highlight the effectiveness of our approach and its potential for advancing virtual screening in drug discovery. Figure 62 illustrates the overall computational workflow employed in the research project.



**Figure 62.** Computational Workflow for predicting molecular activity using MD and 4DES.

## 6.2. Introduction

### 6.2.1. The role of Machine Learning and Artificial Intelligence in Drug Discovery

In recent decades, the field of pharmaceutical chemistry has undergone a significant transformation, driven by the vast amount of chemical, biological, and experimental data available for analysis. The integration of high-performance computing, automated techniques, and advanced data processing methods has enabled a new era in drug discovery, where the design and identification of new drug candidates are supported by computational approaches. Among the most impactful developments in this field is the application of ML, a branch of Artificial Intelligence (AI) that focuses on enabling computers to autonomously learn from data [223].

ML techniques have proven particularly valuable in medicinal chemistry, where they facilitate the connection between molecular features and bioactivity levels, including target affinity and pharmacokinetic properties. By processing available data through suitable ML algorithms, researchers can uncover patterns and insights that traditional methods might overlook. This allows for the prediction of biological activity and other critical pharmacological outcomes, streamlining the drug discovery process, increasing hit rates for promising compounds, and enabling exploration of novel chemical spaces too complex for manual investigation.

ML encompasses four fundamental types: supervised learning utilizes labeled data for specific tasks, unsupervised learning identifies patterns in unlabeled data, semi-supervised learning integrates a small set of labeled data with a larger unlabeled dataset, and reinforcement learning enables an agent to learn through interactions and rewards in an environment. Additionally, various algorithms such as Random Forest (RF), SVM, Neural Networks (NN), and Gradient Boosting (GB) can be applied. RF builds multiple decision trees during training to improve accuracy and robustness. SVM finds a hyperplane in an N-dimensional space to classify data points, NN mimics the brain neural structure to discern complex patterns, and GB builds models sequentially by addressing previous model errors. These methods are selected based on the problem and data type, often combined for optimal performance [223,224].

Today, ML is a cornerstone in drug discovery protocols, accelerating the identification of new therapeutic candidates with greater precision and efficiency. Its integration into

computational chemistry workflows provides unprecedented opportunities to explore and exploit the vast chemical and biological datasets generated in modern research, significantly contributing to the future of medicinal chemistry [225].

However, to develop robust and accurate ML models, access to large quantities of high-quality data and appropriate data representation are essential, ensuring that ML models can process and interpret the data effectively, leading to better predictions of molecular activity and drug potential. The combination of comprehensive datasets and sophisticated molecular representations is thus fundamental to driving further advancements in ML-driven drug discovery.

### **6.2.2. Molecular Features: Molecular Descriptors and ElectroShape Descriptors**

Molecular representation is a crucial aspect of applying computational approaches in pharmaceutical chemistry, as it provides a numerical encoding of the structural and chemical characteristics of molecules. Two key approaches to molecular representation are MD and ElectroShape (ES) Descriptors, each offering unique advantages for modeling the interactions between small molecules and biological targets.

MD are numerical values that summarize various properties of a molecule, transforming its structure into a set of features that can be processed by ML algorithms. These descriptors can be divided into several categories, including topological, geometrical, electronic, and thermodynamic descriptors. For example, MD may quantify properties like molecular weight, hydrogen bond donors/acceptors, surface area, and lipophilicity (e.g., LogP). These features provide a foundational understanding of how molecules might interact with biological environments, and they are essential in establishing Quantitative Structure-Activity Relationships (QSAR). MD have been widely used in cheminformatics and virtual screening, offering a computationally efficient way to represent large libraries of chemical compounds and assess their drug-like properties.

ES Descriptors are an advanced method that integrate shape, chirality, and electrostatics into a unified framework. They build on the Ultrafast Shape Recognition (USR) approach, which computes shape descriptors, by adding partial charge information as a fourth dimension in the case of 4DES, and LogP as a fifth dimension in the case of 5-Dimensional ElectroShape

Descriptors (5DES). This enhancement allows for the integration of electrostatic properties alongside molecular shape, as well as the inclusion of lipophilicity (LogP), providing a more comprehensive and multidimensional representation of molecular similarity.

Additionally, the method incorporates Chiral Shape Recognition (CSR) to differentiate between enantiomers, enhancing the ability to distinguish compounds with similar shapes but different chiral configurations. Specifically, 4DES describe each atom in a molecule using four numbers: three spatial coordinates (x, y, z) and the partial charge (q), while 5DES utilize three spatial coordinates (x, y, z), the partial charge (q), and the LogP (p), situating the molecule in a four-dimensional space (R4) and five-dimensional space (R5), respectively. The similarity between molecules is then calculated by comparing these descriptors, enabling faster and more precise identification of molecules with similar binding properties.

In practice, both MD and ES Descriptors play a complementary role. MD provide a computationally efficient method for representing a broad range of chemical properties, whether in 2D, 3D, or 4D form. Meanwhile, ES offers deeper insights into the electronic and spatial attributes of molecules, critical for predicting specific interactions with biological targets. This combination of approaches allows ML models to capture both general chemical diversity and detailed bioactivity information, leading to more accurate predictions and better-informed decisions in the drug discovery pipeline.

### **6.2.3. Useful Benchmarking sets**

The quality and type of data used are critical factors when developing ML models. Utilizing established datasets for training and validation, rather than synthetic or augmented data, ensures that the training set is representative, exhaustive, and diverse.

The performance of various ML models is typically evaluated through ligand enrichment using benchmarking sets, measuring the ability to rank true ligands highly in a list of potential candidates against a large pool of decoys, which are assumed inactive and unlikely to bind to the target. This combination of true ligands and their associated decoys constitutes the benchmarking set [226-235].

Benchmarking sets can be classified into two major types according to their initial designing purposes, the Structure Based Virtual Screening (SBVS) specific and the Ligand Based Virtual Screening (LBVS) specific. Datasets such as Directory of Useful Decoys (DUD) [236] and its recent DUD-Enhanced (DUD-E) [237], Virtual Decoy Sets (VDS) [238], G Protein-Coupled

Receptors (GPCRs) Ligand Library (GLL) and GPCRs Decoy Database (GDD) [239], Demanding Evaluation Kits for Objective *in Silico* Screening (DEKOIS) [240] and DEKOIS 2.0 [241], Nuclear Receptors Ligands and Structures Benchmarking DataBase (NRLiSt BDB) belong to SBVS-specific benchmarking sets. By contrast, DUD LIB VS 1.0 [242], database of *reproducible virtual screens* (REPROVIS-DB) [243], Maximum Unbiased Validation (MUV) database [221], and LIT-PubChemBioActivity (LIT-PCBA) database [222] are specifically designed for the purpose of LBVS. To date, DUD and DUD-E have been intensively employed as gold standard data sets among the community [212,213,218,244-248], while much fewer citations of MUV [213,215,218,219] have been reported.

### 6.2.3.1. DUD and DUD-E datasets

The DUD dataset [236] was created to address the issue of artificial enrichment in SBVS by ensuring that decoys have similar physicochemical properties but distinct 2D topology compared to ligands. DUD contains 95,316 decoys and 2,950 ligands, targeting 40 proteins across six categories: nuclear hormone receptors, kinases, serine proteases, metalloenzymes, folate enzymes, and other enzymes. The decoys were sourced from ZINC's "drug-like" subset, using a method based on the Tanimoto coefficient (Tc) and 32 calculated physicochemical properties, with a focus on five key properties—molecular weight (MW), hydrogen bond acceptors (HBAs), hydrogen bond donors (HBDs), rotatable bonds (RBs), and LogP—relevant to drug-likeness based on Lipinski's rule of five. For each ligand, 36 compounds with the most similar physicochemical properties were selected as decoys. DUD decoys accurately match ligands in their physical property distributions, thus reducing artificial enrichment bias. Publicly available at <http://dud.docking.org/>, DUD became the first bias-corrected, SBVS-specific benchmarking set and quickly became a standard for structural model validation and software evaluation [249-251].

However, to mitigate "analogue bias" [252], lead-like and reduced-graph filters were applied to DUD, resulting in the "DUD Clustering" dataset, which encourages chemical diversity and improves scaffold hopping evaluations. This refined set is available at <http://dud.docking.org/clusters/> and is recommended to be used in conjunction with the original DUD decoys for ligand enrichment studies.

The enhanced version of DUD, known as DUD-E <https://dude.docking.org/> [237], was released to further address the limitations of the original set. DUD-E includes 22,886 active compounds across 102 diverse protein targets and provides approximately 50 decoys for each active compound. The decoys are similarly selected to share physicochemical properties with the active compounds while maintaining dissimilar 2D topologies, drawn from ZINC's database. DUD-E is primarily used in classification tasks to benchmark the performance of molecular docking programs, particularly in terms of their ability to rank active compounds higher than decoys. Due to its larger scale and improved design, DUD-E is considered a more comprehensive and reliable benchmarking set for modern virtual screening studies.

### 6.2.3.2. MUV dataset

The MUV set is a carefully designed dataset aimed at providing unbiased validation for virtual screening methods, particularly LBVS. Its ligands and decoys are sourced from Primary and Confirmatory BioAssays (PCBioAssay) in PubChem [253]. Specifically, inactive compounds from primary bioassays against a target were used to build the "Potential Decoys" data set, while compounds that were confirmed active through both primary and confirmatory bioassays formed the "Potential Actives" dataset. To further enhance reliability and minimize false positive, the "Potential Actives" dataset was refined using an assay artifacts filter, which includes the "Hill slope filter", the "frequency of hits filter", and the "autofluorescence and luciferase inhibition filter". A major strength of the MUV set is its exclusive use of experimentally confirmed inactives as decoys, a significant advantage over other benchmarking datasets. Each target's final benchmark comprises 30 active compounds and 15,000 decoys. Through the analysis of 17 MUV benchmarking sets (MUV codes per each target are shown in Table 18), it was demonstrated that this approach effectively reduces both artificial enrichment and analogue bias. The datasets [221], along with a tool for generating MUV decoys, are publicly available at <http://www.pharmchem.tu-bs.de/lehre/baumann/MUV.html>.

**Table 18.** Details of the MUV datasets, including the MUV ID, target name, and type of interaction associated with each target.

MUV ID	Target name	Interaction
466	S1P1 receptor*	Agonists

548	PKA*	Inhibitors
600	SF1*	Inhibitors
644	Rho Kinase2	Inhibitors
652	HIV RT-RNase*	Inhibitors
689	Eph rec. A4*	Inhibitors
692	SF1*	Agonists
712	HSP90*	Inhibitors
713	ER- $\alpha$ -coact. bind.*	Inhibitors
733	ER- $\beta$ -coact. bind.*	Inhibitors
737	ER- $\alpha$ -coact. bind.*	Potentiators
810	FAK*	Inhibitors
832	Cathepsin G	Inhibitors
846	FXIa*	Inhibitors
852	FXIIa*	Inhibitors
858	D1 receptor*	Allosteric modulators
859	M1 receptor*	Allosteric inhibitors

\*Abbreviations: Sphingosine-1-Phosphate receptor 1 (S1P1 receptor); Protein Kinase A (PKA); nuclear receptor Steroidogenic Factor 1 (SF1); Human Immunodeficiency Virus Reverse Transcriptase RNase (HIV RT-RNase); Eph receptor A4 (Eph rec. A4); Heat Shock Protein 90 (HSP90); Estrogen Receptor- $\alpha$ -coactivator binding (ER- $\alpha$ -coact. bind); Estrogen Receptor- $\beta$ -coactivator binding (ER- $\beta$ -coact. bind); Focal Adhesion Kinase (FAK); Coagulation Factor XIa (FXIa); Coagulation Factor XIIa (FXIIa); M1 muscarinic receptor (M1 receptor); D1 dopamin receptor (D1 receptor).

When compared to the MUV dataset, DUD exhibits significant biases. Many DUD data sets show "clumping" in descriptor space, making the retrieval of active compounds relatively easy, even with simple descriptors, indicating a substantial benchmark bias. Additionally, DUD suffers from high analogue bias, with the average number of compounds per scaffold class being 4.56, compared to 1.16 in MUV. While this is less problematic for docking methods, it critically impacts LBVS. Furthermore, despite its design to minimize separation between actives and decoys, considerable separation is still observed, likely due to the strict Tanimoto dissimilarity criterion, leading to poor embedding of actives within decoys. In some cases, up to 60% of actives are not well surrounded by decoys. In contrast, MUV avoids these issues by using experimentally confirmed inactive decoys, eliminating the need for dissimilarity criteria and providing a more rigorous and unbiased validation set for LBVS, making it superior to DUD in this context.

### **6.2.3.3. LIT-PCBA dataset**

The LIT-PCBA dataset [222] represents a novel generation of virtual screening benchmarking datasets, specifically aimed at advancing ML applications in the prediction of bioactivity.

Constructed from dose-response bioassays available in PubChem, it features unambiguously defined active and inactive compounds. A meticulous examination of the metadata allowed for the removal of assay artifacts, frequent hitters, and false positives, resulting in a high-quality dataset. LIT-PCBA encompasses 15 target sets, covering a diverse array of ligands and target proteins. Preliminary virtual screening attempts utilizing state-of-the-art methods such as 2D similarity, 3D shape matching, and docking have revealed the dataset's challenging nature, particularly due to the absence of potency distribution biases among labeled active compounds. Furthermore, a recently described unbiasing procedure was applied to optimize the distribution of training and validation compounds for ML applications. This unique challenge is expected to enhance the evaluation of modern artificial intelligence methods in structure-based virtual screening scenarios. The LIT-PCBA dataset is freely accessible at <https://drugdesign.unistra.fr/LIT-PCBA/>.



### 6.3. Aim of the project

ML has emerged as a transformative technology in various fields, including pharmaceuticals and biomedical research, by enabling the analysis of large datasets to make predictions and uncover insights that were previously unattainable. As traditional methods of drug discovery become increasingly time-consuming and resource-intensive, ML offers innovative solutions to enhance the efficiency and effectiveness of virtual screening processes. This project aims to develop a robust ML model that effectively discriminates between active and inactive compounds toward specific biological targets, thereby streamlining the drug discovery process.

A critical challenge in applying ML to drug discovery is learning useful molecular representations. In our approach, we utilize a combination of MD and 4DES to represent small molecules effectively. MD are computed using the MOLDESTO software, capturing various physicochemical properties vital for understanding molecular behavior. In contrast, 4DES provide an advanced representation that incorporates not only the three-dimensional spatial arrangement of molecules but also their electronic characteristics. This multifaceted descriptor framework is expected to yield a richer and more nuanced understanding of molecular differences, which is essential for accurately distinguishing between bioactive and non-bioactive compounds.

Our project primarily focuses on the MUV dataset, which is known for minimizing analogue bias and artificial enrichment, thus providing a reliable foundation for training, and validating our ML models. The MUV dataset includes assay data for 17 diverse biological targets, enabling us to develop a more generalized model capable of performing well across various contexts. To assess the performance of our ML model, SVM approach is implemented, leveraging 5-fold cross-validation as a means of evaluation. Performance metrics such as AUC and the 1%EF are employed to quantify the model's effectiveness in classifying compounds. Additionally, a 3-fold cross-validation is performed by stratifying the MUV dataset based on target numerical codes and FASTA sequence identity. This methodology aims to test the model's robustness and its ability to generalize across different biological contexts. We further evaluate the model's performance using the LIT-PCBA dataset, which presents its own set of challenges and opportunities for enhancing predictive accuracy.

In conclusion, this project aims to contribute to the growing field of computational drug discovery by developing an ML model that not only demonstrates high performance in predicting molecular activity but also provides insights into the underlying features that influence bioactivity. By leveraging advanced descriptor methodologies and rigorous validation techniques, we hope to establish a framework that enhances virtual screening capabilities, ultimately aiding in the identification of promising therapeutic candidates.

## 6.4. Results and Discussion

### 6.4.1. Molecular and ElectroShape Descriptors for MUV target sets

As the first step of the work, the calculation of MD and ES Descriptors was performed for the 17 biological targets in the MUV dataset. The goal was to represent each molecule's structural and electronic properties comprehensively, which would serve as input features for the subsequent ML model to classify compounds as active or inactive (binary classification: 1 for active, 0 for inactive).

The MD were generated using MOLDESTO [108], an in-house software developed by the research group at the University of Palermo. MOLDESTO is designed to compute a broad range of molecular properties, calculating approximately 1,000 descriptors spanning 1D, 2D, and 3D molecular features for each molecule (detailed in List S1 of the Supplementary Material; see Materials and Methods section, paragraph 4.1.1.1).

In parallel, ES descriptors were calculated using EShape, Oxford Drug Design's proprietary software for ES descriptor calculation [220]. Since the lipophilic properties (e.g., LogP) are already included in the MD set, we opted not to calculate 5DES and instead used 4DES. This approach still allows for a sophisticated representation of a molecule's spatial configuration and electrostatic properties. The 4DES methodology also accounts for chirality and differentiates between enantiomers, providing valuable insights into subtle differences that may influence molecular interactions with target proteins.

Together, these two sets of descriptors provide complementary representation of input molecules. While the MD are computationally efficient in capturing a broad range of molecular features, the 4DES descriptors offer a more detailed insight into 3D shape and electrostatic complementarity of the molecules, which are critical for predicting biological activity.

Each molecule from the MUV dataset was thus characterized by these descriptors, setting up the dataset for the ML classification task, where compounds labeled as active (Class 1) or inactive (Class 0) based on their biological assay results. This dataset forms the foundation for training and evaluating the ML models developed in this study.

### 6.4.2. Machine Learning Model Development

In this section, we describe the development of the ML model aimed at classifying molecular compounds as active or inactive based on MD and 4DES descriptors.

The SVM model was employed due to its efficiency in high-dimensional datasets and its ability to handle binary classification problems. To evaluate the performance of the ML model, we implemented a stratified 5-fold cross-validation scheme. Stratification ensures that each fold maintains a balanced proportion of active (Class 1) and inactive (Class 0) compounds, preventing bias in the model evaluation. The model is trained on 80% of the data and tested on the remaining 20%, repeated across five different data splits to obtain reliable performance metrics.

The parameters utilized for the ML model performance evaluation are Receiver Operating Characteristic - Area Under the Curve (ROC-AUC), and the EF. The ROC curve plots the True Positive Rate (TPR) against the False Positive Rate (FPR), while the AUC provides a summary of the model's classification capability, with higher values indicating better performance. In addition, the EF is a crucial metric in virtual screening applications, as it evaluates the model's ability to prioritize active compounds in the top-ranked subset. EF is computed as follows:

$$EF_{\alpha} = \frac{NTB_{\alpha}}{NTB_{total} * \alpha}$$

where  $NTB_{\alpha}$  is the Number of True Binders among top  $\alpha$  percentile of ranked candidates (1%, 5%, 10%) based on predicted binding scores and  $NTB_{total}$  is the total Number of True Binders in the database. Here, EF was computed for the top 1% of predictions (1%EF), which gives insight into the model's effectiveness at retrieving actives.

SHAP method is used to assess the contribution of each feature to the final predictions.

### 6.4.3. Preliminary Results on MUV set – I ROUND

Once the MD and 4DES descriptors for MUV dataset were computed and the SVM model was developed, the molecules from the MUV dataset were submitted to the ML model for preliminary testing. This initial round aimed to evaluate the model's ability to discriminate between active and inactive compounds. The performance of the ML model in terms of AUC and 1%EF was evaluated across 17 MUV targets using three different methods: MD, 4DES, and the combined approach of MD+4DES.

In this first round, the model yielded an average AUC of  $0.763 \pm 0.099$  when using the combined MD and 4DES representations. When using MD alone, the model achieved an average AUC of  $0.759 \pm 0.099$ , while using 4DES alone resulted in a lower average AUC ( $0.635 \pm 0.113$ ). MD generally performed well across the targets, with AUC values mostly (14/17 targets) above 0.7, indicating good model performance. The highest performance was observed for MUV\_852 target, where the MD method achieved an AUC of  $0.923 \pm 0.025$ . This suggests that MD descriptors are particularly effective in capturing the relevant chemical information for this target. However, for certain targets like MUV\_600 (AUC:  $0.590 \pm 0.075$ ) and MUV\_692 (AUC:  $0.550 \pm 0.098$ ), MD underperformed, suggesting that molecular descriptors alone may not provide sufficient discriminatory power for these targets.

Across all targets, the combined MD+4DES approach yields AUC values comparable to, or slightly better than, those obtained using either MD or 4DES individually. However, the difference between MD and MD+4DES is not statistically significant for most, if not all, targets. Table 19 provides a detailed comparison of the AUC values for each MUV target when using MD, 4DES, and the combined (MD+4DES) approach.

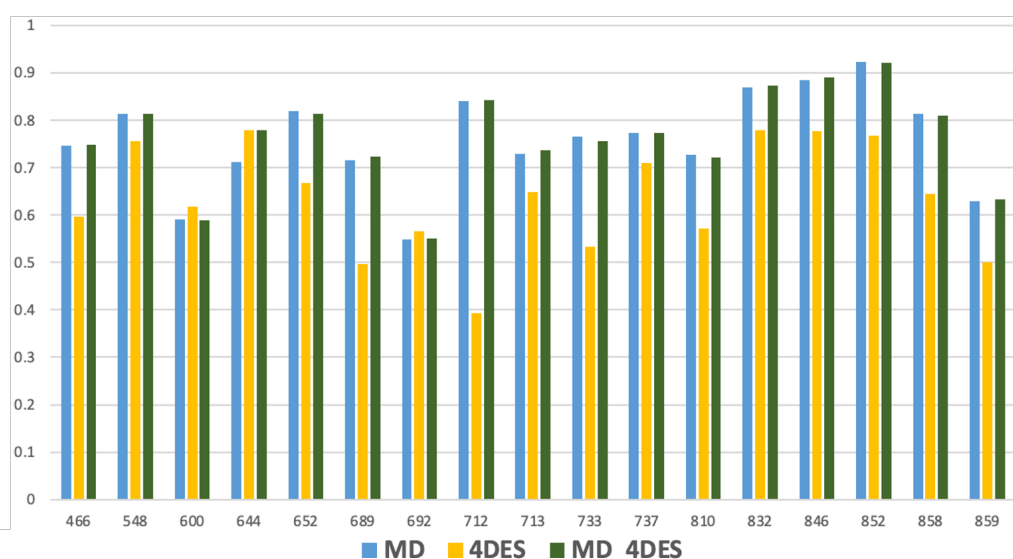
**Table 19.** Summary of the AUC values for all 17 MUV targets, comparing the performance of models built using MD, 4DES, and the combined MD+4DES descriptor sets.

MUV ID	Target name	MD	4DES	MD+4DES
MUV_466	S1P1 receptor*	$0.747 \pm 0.053$	$0.597 \pm 0.105$	$0.749 \pm 0.050$
MUV_548	PKA*	$0.814 \pm 0.085$	$0.756 \pm 0.102$	$0.814 \pm 0.086$
MUV_600	SF1*	$0.590 \pm 0.075$	$0.617 \pm 0.076$	$0.588 \pm 0.079$
MUV_644	Rho Kinase2	$0.712 \pm 0.127$	$0.778 \pm 0.045$	$0.714 \pm 0.119$
MUV_652	HIV RT-RNase*	$0.819 \pm 0.080$	$0.668 \pm 0.103$	$0.813 \pm 0.083$
MUV_689	Eph rec. A4*	$0.716 \pm 0.024$	$0.496 \pm 0.113$	$0.724 \pm 0.022$
MUV_692	SF1*	$0.550 \pm 0.098$	$0.565 \pm 0.037$	$0.548 \pm 0.105$
MUV_712	HSP90*	$0.841 \pm 0.065$	$0.392 \pm 0.105$	$0.842 \pm 0.063$
MUV_713	ER- $\alpha$ -coact. bind.*	$0.728 \pm 0.070$	$0.649 \pm 0.124$	$0.736 \pm 0.065$
MUV_733	ER- $\beta$ -coact. bind.*	$0.765 \pm 0.162$	$0.533 \pm 0.048$	$0.756 \pm 0.163$
MUV_737	ER- $\alpha$ -coact. bind.*	$0.774 \pm 0.088$	$0.710 \pm 0.081$	$0.774 \pm 0.090$
MUV_810	FAK*	$0.727 \pm 0.108$	$0.571 \pm 0.129$	$0.721 \pm 0.112$
MUV_832	Cathepsin G	$0.869 \pm 0.060$	$0.778 \pm 0.124$	$0.872 \pm 0.052$

<b>MUV_846</b>	FXIa*	0.885 ± 0.057	0.776 ± 0.066	0.891 ± 0.056
<b>MUV_852</b>	FXIIa*	0.923 ± 0.025	0.767 ± 0.066	0.921 ± 0.029
<b>MUV_858</b>	D1 receptor*	0.813 ± 0.050	0.645 ± 0.148	0.810 ± 0.057
<b>MUV_859</b>	M1 receptor*	0.630 ± 0.069	0.501 ± 0.072	0.633 ± 0.072
<b>AVERAGE</b>		0.759 ± 0.099	0.635 ± 0.113	0.763 ± 0.099

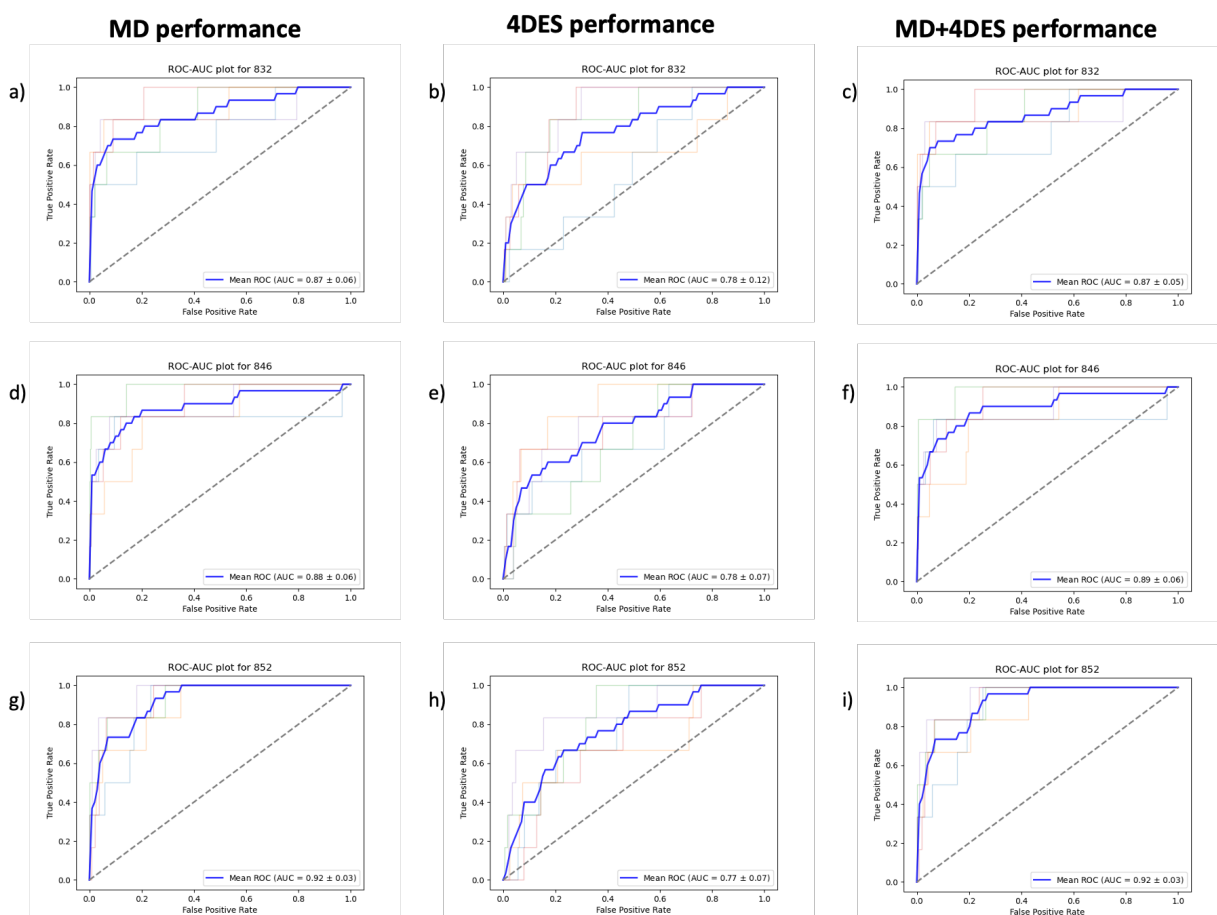
\*Abbreviations: Sphingosine-1-Phosphate receptor 1 (S1P1 receptor); Protein Kinase A (PKA); nuclear receptor Steroidogenic Factor 1 (SF1); Human Immunodeficiency Virus Reverse Transcriptase RNase (HIV RT-RNase); Eph receptor A4 (Eph rec. A4); Heat Shock Protein 90 (HSP90); Estrogen Receptor- $\alpha$ -coactivator binding (ER- $\alpha$ -coact. bind); Estrogen Receptor- $\beta$ -coactivator binding (ER- $\beta$ -coact. bind); Focal Adhesion Kinase (FAK); Coagulation Factor XIa (FXIa); Coagulation Factor XIIa (FXIIa); M1 muscarinic receptor (M1 receptor); D1 dopamin receptor (D1 receptor).

This result, that strongly emphasizes the importance of combining features for targets, are graphically represented in Figure 63, where the histogram illustrates the variation in AUC values across the 17 MUV targets for the three descriptor methods: MD (blue), 4DES (yellow), and the combination of MD+4DES (green). The color-coding allows for a clear visual comparison of how each method influences model performance across different target.



**Figure 63.** AUC values for each of the 17 MUV targets across the three descriptor methods: MD (blue), 4DES (yellow), and MD+4DES (green). The histogram illustrates the variation in model performance for each target.

ROC-AUC for the best-performing targets is shown in Figure 64 (for the other targets, see Figures S28-S32 of the Supplementary Material). Specifically, this figure illustrates the ROC-AUC curves for MUV\_832, MUV\_846, and MUV\_852, comparing the performance across the MD, 4DES, and combined MD+4DES descriptor sets.



**Figure 64.** (a) ROC-AUC for target MUV\_832 when using the MD approach; (b) ROC-AUC for target MUV\_832 when using the 4DES approach; (c) ROC-AUC for target MUV\_832 when using the MD+4DES combined approach; (d) ROC-AUC for target MUV\_846 when using the MD approach; (e) ROC-AUC for target MUV\_846 when using 4DES approach; (f) ROC-AUC for target MUV\_846 when the using the MD+4DES combined approach; (g) ROC-AUC for target MUV\_852 when using the MD approach; (h) ROC-AUC for target MUV\_852 when using the 4DES approach; (i) ROC-AUC for target MUV\_852 when using the MD+4DES combined approach.

The EF analysis, shown in Table 20, complements the AUC evaluation by providing insight into the model's ability to prioritize active compounds in the top-ranking predictions. While AUC reflects the overall discriminative power of the models, EF focuses on the early retrieval of active compounds, a critical metric in virtual screening scenarios. Across the 17 MUV targets, MD, and the MD+4DES combination consistently outperform the 4DES descriptors alone in most cases, highlighting the added value of incorporating molecular descriptors into the model. Notably, targets such as MUV\_832, MUV\_846, and MUV\_852 exhibit the highest EF values, particularly using the MD and MD+4DES models. For example, MUV\_846 reaches an impressive EF value of  $51.72 \pm 15.84$  with both the MD and the combined methods, making it the highest-performing target in the study. Similarly, MUV\_832 shows an EF of  $45.25 \pm 12.09$

using the same approaches. The combination of MD and 4DES descriptors demonstrates strong results, with some targets benefiting from the integration of both descriptor sets. In particular, for targets such as MUV\_548 and MUV\_737, the combined approach outperforms individual descriptor methods.

Conversely, the 4DES method alone often struggles to achieve competitive EF values, particularly for targets like MUV\_466, MUV\_712, and MUV\_713, MUV\_733, MUV\_737, MUV\_858, and MUV\_859, where zero EF values are observed. This suggests that the 4DES descriptors, when used in isolation, are less effective in identifying MUV-active compounds. However, MD alone also yielded zero EF values for two targets (MUV\_692 and MUV\_859). Overall, the combined MD+4DES method consistently perform well, maintaining higher EF values across most targets. This demonstrates that relying on either MD or 4DES alone increases the risk of suboptimal performance, whereas their combination enhances model effectiveness.

**Table 20.** Summary of the 1%EF values for all 17 MUV targets, comparing the performance of models built using MD, 4DES, and the combined MD+4DES descriptor sets.

MUV ID	Target name	MD	4DES	MD+4DES
MUV_466	S1P1 receptor*	16.16 ± 10.22	0	16.16 ± 10.22
MUV_548	PKA*	16.16 ± 10.22	9.70 ± 12.93	19.39 ± 15.84
MUV_600	SF1*	12.93 ± 6.47	9.70 ± 7.92	12.93 ± 6.47
MUV_644	Rho Kinase2	25.86 ± 12.93	6.47 ± 12.93	25.86 ± 12.93
MUV_652	HIV RT-RNase*	19.391 ± 6.47	6.47 ± 7.92	19.39 ± 6.47
MUV_689	Eph rec. A4*	19.39 ± 12.09	6.47 ± 7.92	19.39 ± 12.09
MUV_692	SF1*	0	3.23 ± 6.46	0
MUV_712	HSP90*	25.85 ± 16.48	0	25.85 ± 16.48
MUV_713	ER-α-coact. bind.*	22.63 ± 7.92	0	22.63 ± 7.92
MUV_733	ER-β-coact. bind.*	9.70 ± 12.93	0	9.70 ± 12.93
MUV_737	ER-α-coact. bind.*	9.70 ± 7.92	0	12.93 ± 6.47
MUV_810	FAK*	3.23 ± 6.47	6.47 ± 7.92	3.23 ± 6.47
MUV_832	Cathepsin G	45.25 ± 12.09	19.39 ± 12.09	45.25 ± 12.09
MUV_846	FXIa*	51.72 ± 15.84	9.70 ± 7.92	51.72 ± 15.84
MUV_852	FXIIa*	32.32 ± 10.22	3.23 ± 6.47	32.32 ± 10.22



<b>MUV_858</b>	D1 receptor*	9.70 ± 12.93	0	9.70 ± 12.93
<b>MUV_859</b>	M1 receptor*	0	0	0

\*Abbreviations: Sphingosine-1-Phosphate receptor 1 (S1P1 receptor); Protein Kinase A (PKA); nuclear receptor Steroidogenic Factor 1 (SF1); Human Immunodeficiency Virus Reverse Transcriptase RNase (HIV RT-RNase); Eph receptor A4 (Eph rec. A4); Heat Shock Protein 90 (HSP90); Estrogen Receptor- $\alpha$ -coactivator binding (ER- $\alpha$ -coact. bind); Estrogen Receptor- $\beta$ -coactivator binding (ER- $\beta$ -coact. bind); Focal Adhesion Kinase (FAK); Coagulation Factor XIa (FXIa); Coagulation Factor XIIa (FXIIa); M1 muscarinic receptor (M1 receptor); D1 dopamin receptor (D1 receptor).

After analyzing the model performance through AUC and 1%EF for the 17 targets of the MUV dataset, it is essential to further refine our models. Since MD computes about 1,000 descriptors per molecule, which is a high dimensional representation for molecules, we employed the SHAP method [254] to reduce the dimensionality and identify the 50 most significant descriptors for each target in the MD runs.

This approach not only enhances our understanding of the importance of each descriptor in the model's decision-making process but also facilitates dimensionality reduction, thereby improving the interpretability and effectiveness of our predictive models when testing large dataset of compounds in virtual screening projects. By leveraging the insights provided by SHAP, we can focus on a more relevant set of descriptors, ultimately optimizing the overall performance of our models.

Notably, SHAP was not applied to the 4DES descriptors, as the number of ES descriptors is already small (15).

Consequently, we obtained new matrices for each target with a reduced set of descriptors, consisting of 50 MD and 15 4DES. We utilized these new matrices to rerun our model in a second round, allowing us to assess the impact of the selected descriptors on model performance more effectively. The Top50 descriptors list for each target is reported in Supplementary Material S1, Table S10

#### **6.4.4. Results on MUV set – II ROUND**

In this section, we present the results of the second round of testing, where the MUV dataset was re-evaluated using the ML model after selecting the top 50 descriptors based on SHAP analysis of the MD and MD+4DES runs from the first round (for top50 descriptors list for each target see Supplementary Material S1, Table S10). This approach aims to enhance model performance by focusing on the most relevant features for each target. In the second round of analysis, the AUC values reveal that the selection of the top 50 MDs has led to better results

across the 17 MUV targets. Table 21 makes it possible to appreciate the difference between AUC values across rounds I and II, highlighting the overall improvement in predictive accuracy as a result of the refined feature selection process.

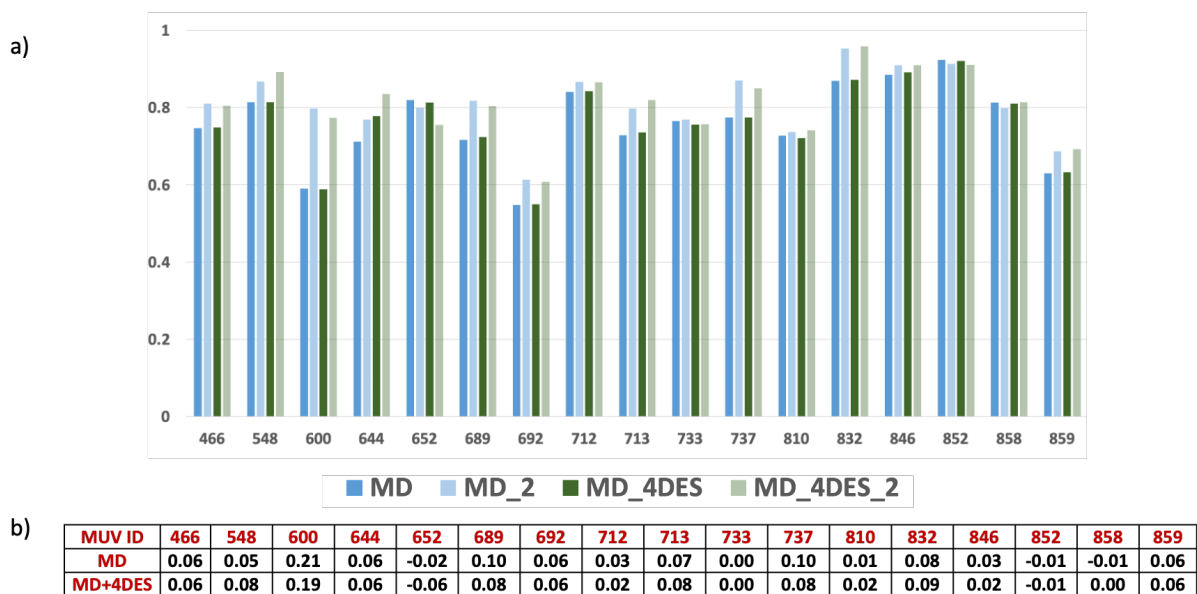
**Table 21.** Comparison of AUC values for MD, MD+4DES, and 4DES methods across rounds I and II. The second round uses a refined feature selection process based on the top 50 MD determined by SHAP analysis, showing overall improvement in average AUC scores compared to the first round, with minimal difference between the MD-only and MD+4DES methods.

MUV ID	MD (I ROUND)	MD (II ROUND)	MD+4DES (I ROUND)	MD+4DES (II ROUND)	4DES (I and II ROUND)
MUV_466	0.747 ± 0.053	0.810 ± 0.071	0.749 ± 0.050	0.805 ± 0.071	0.597 ± 0.105
MUV_548	0.814 ± 0.085	0.867 ± 0.080	0.814 ± 0.086	0.892 ± 0.054	0.756 ± 0.102
MUV_600	0.590 ± 0.075	0.797 ± 0.083	0.588 ± 0.079	0.773 ± 0.055	0.617 ± 0.076
MUV_644	0.712 ± 0.127	0.769 ± 0.111	0.714 ± 0.119	0.835 ± 0.062	0.778 ± 0.045
MUV_652	0.819 ± 0.080	0.799 ± 0.092	0.813 ± 0.083	0.755 ± 0.082	0.668 ± 0.103
MUV_689	0.716 ± 0.024	0.818 ± 0.056	0.724 ± 0.022	0.804 ± 0.027	0.496 ± 0.113
MUV_692	0.550 ± 0.098	0.613 ± 0.138	0.548 ± 0.105	0.608 ± 0.134	0.565 ± 0.037
MUV_712	0.841 ± 0.065	0.866 ± 0.065	0.842 ± 0.063	0.865 ± 0.047	0.392 ± 0.105
MUV_713	0.728 ± 0.070	0.797 ± 0.039	0.736 ± 0.065	0.819 ± 0.041	0.649 ± 0.124
MUV_733	0.765 ± 0.162	0.769 ± 0.126	0.756 ± 0.163	0.757 ± 0.120	0.533 ± 0.048
MUV_737	0.774 ± 0.088	0.870 ± 0.048	0.774 ± 0.090	0.850 ± 0.069	0.710 ± 0.081
MUV_810	0.727 ± 0.108	0.737 ± 0.089	0.721 ± 0.112	0.741 ± 0.116	0.571 ± 0.129
MUV_832	0.869 ± 0.060	0.953 ± 0.028	0.872 ± 0.052	0.958 ± 0.032	0.778 ± 0.124
MUV_846	0.885 ± 0.057	0.910 ± 0.052	0.891 ± 0.056	0.910 ± 0.044	0.776 ± 0.066
MUV_852	0.923 ± 0.025	0.913 ± 0.045	0.921 ± 0.029	0.911 ± 0.057	0.767 ± 0.066
MUV_858	0.813 ± 0.050	0.798 ± 0.056	0.810 ± 0.057	0.814 ± 0.065	0.645 ± 0.148
MUV_859	0.630 ± 0.069	0.687 ± 0.102	0.633 ± 0.072	0.692 ± 0.098	0.501 ± 0.072
<b>AVERAGE</b>	0.759 ± 0.099	0.810 ± 0.082	0.763 ± 0.099	0.811 ± 0.084	0.635 ± 0.113

The comparison of AUC scores between the first and second rounds reveals significant improvements in overall model performance. In the first round, the average AUC for the model using MD descriptors alone was 0.759, which increased to 0.810. Similarly, the MD+4DES method showed a rise from 0.763 to 0.811, indicating that selecting the top 50 MD descriptors effectively captured more relevant molecular features, thereby enhancing

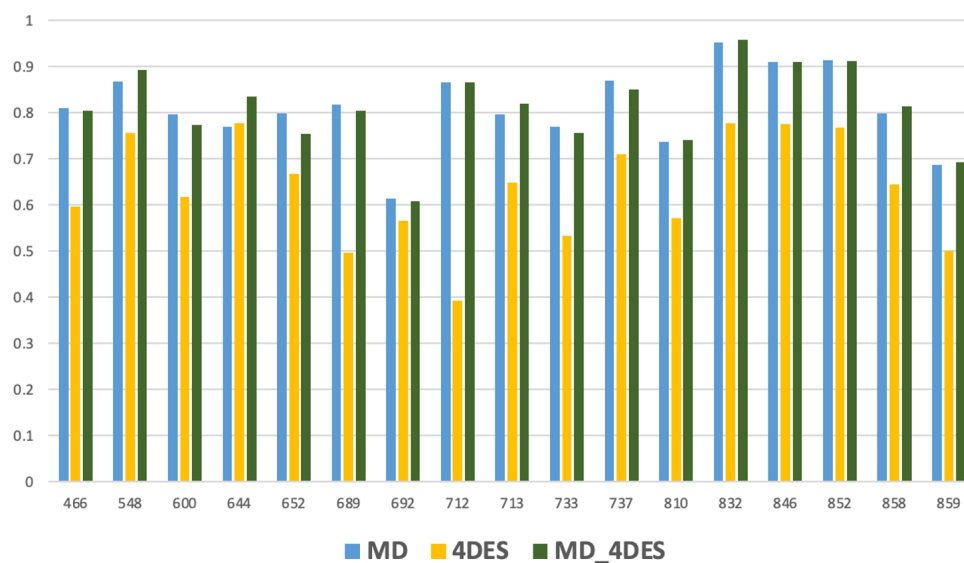
predictive capability. Target-specific analysis further underscores these advancements, with notable improvements across various MUV targets. For instance, MUV\_466 demonstrated a significant increase from 0.747 to 0.810 (MD) and from 0.749 to 0.805 (MD+4DES), while MUV\_548 rose from 0.814 to 0.867 (MD) and from 0.814 to 0.892 (MD+4DES), indicating improved predictive performance. Targets such as MUV\_832 and MUV\_846 also showcased substantial gains, with AUCs increasing from 0.869 to 0.953 and from 0.885 to 0.910 for MD, respectively, and from 0.872 to 0.958 and from 0.891 to 0.910 for MD+4DES, respectively. Conversely, some targets exhibited mixed results; MUV\_600 improved dramatically from 0.590 to 0.797 (MD) and from 0.588 to 0.773 (MD+4DES), as well as MUV\_644, MUV\_689, MUV\_692, MUV\_713, and MUV\_859. Notably, MUV\_737 showed significant improvement, achieving AUC values of 0.870 (MD) and 0.850 (MD+4DES), from 0.774 in both cases. Meanwhile, MUV\_712, MUV\_733, MUV\_810 exhibited slight enhancements, whereas MUV\_652, MUV\_852, and MUV\_858 represent the only instances of decreased performance in the study, albeit with minimal decline.

The accompanying histograms, in Figure 65, visually illustrate these results, highlighting the gaps in AUC scores for each target and effectively demonstrating the improvements made from the first to the second round. While Figure 66 illustrates the variation in AUC values across the 17 MUV targets for the three descriptor methods within the II Round.



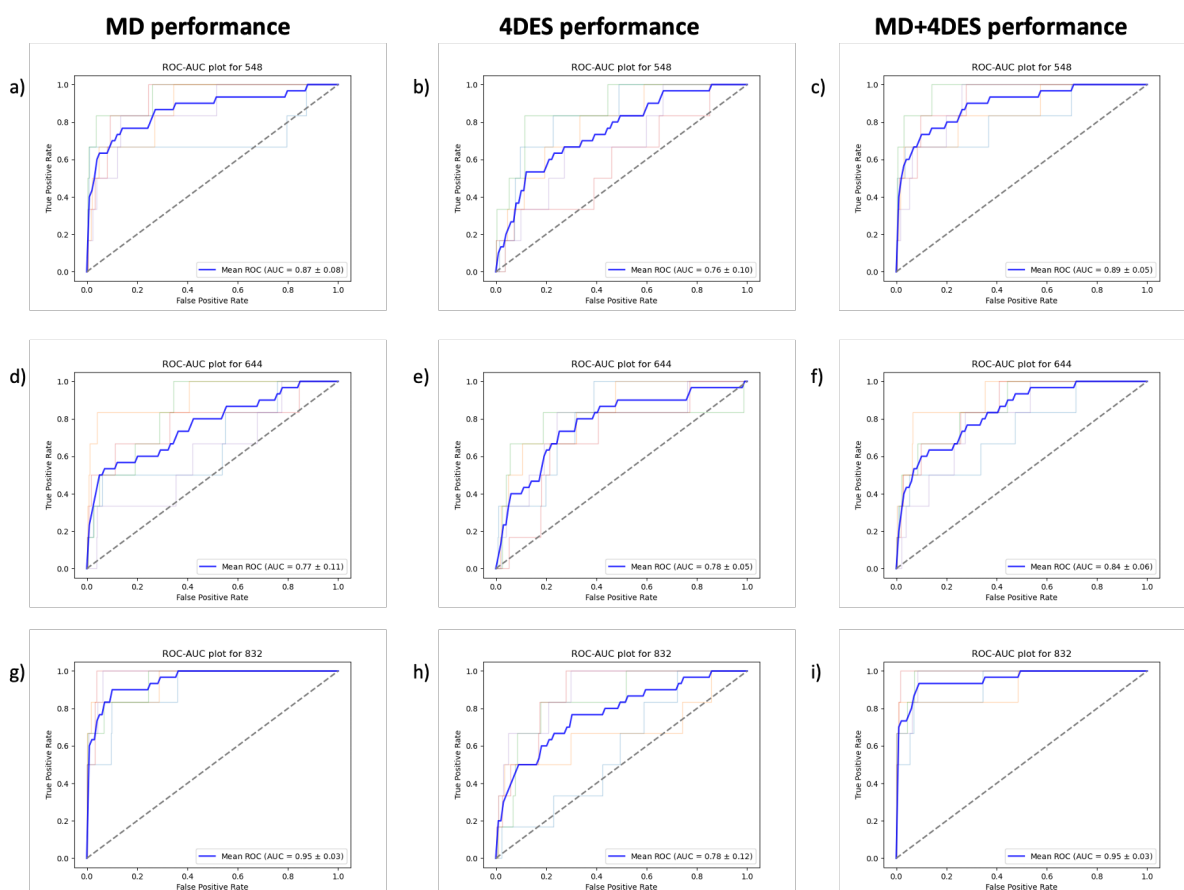
**Figure 65.** (a) Histograms showing the AUC values for MD and MD+4DES across both rounds of analysis; MD and MD+4DES in the first round are represented in blue and green, respectively, while

MD and MD+4DES in the second round in light blue and light green, respectively; **(b)** table highlighting the gaps in performance between the two rounds for MD and MD+4DES.



**Figure 66.** AUC values for each of the 17 MUV targets across the three descriptor methods: MD (blue), 4DES (yellow, where values are identical to the first round), and MD+4DES (green). The histogram illustrates the variation in model performance for each target.

ROC-AUC for the best-performing targets are shown in Figure 67 (for the other targets, see Figures S33-S37 of the Supplementary Material). Specifically, this figure illustrates the ROC-AUC for MUV\_548, MUV\_644, and MUV\_832, comparing the performance across the MD, 4DES, and combined MD+4DES descriptor sets.



**Figure 67.** (a) ROC-AUC for target MUV\_548 when using the MD approach; (b) ROC-AUC for target MUV\_548 when using the 4DES approach; (c) ROC-AUC curve for target MUV\_548 when using the MD+4DES combined approach; (d) ROC-AUC curve for target MUV\_644 when using the MD approach; (e) ROC-AUC curve for target MUV\_644 when using 4DES approach; (f) ROC-AUC curve for target MUV\_644 when using the MD+4DES combined approach; (g) ROC-AUC curve for target MUV\_832 when using the MD approach; (h) ROC-AUC curve for target MUV\_832 when using the 4DES approach; (i) ROC-AUC curve for target MUV\_832 when using the MD+4DES combined approach.

Finally, Table 22 complements the AUC evaluation by providing the 1%EF analysis for each target within the MUV dataset across the Rounds I and II.

**Table 22.** Comparison of 1%EF values for MD, MD+4DES and 4DES methods across Rounds I and II.

MUV ID	MD (I ROUND)	MD (II ROUND)	MD+4DES (I ROUND)	MD+4DES (II ROUND)	4DES (I and II ROUND)
MUV_466	16.16 ± 10.22	19.39 ± 12.09	16.16 ± 10.22	12.93 ± 12.09	0
MUV_548	16.16 ± 10.22	35.56 ± 18.85	19.39 ± 15.84	38.79 ± 16.48	9.70 ± 12.93
MUV_600	12.93 ± 6.47	16.16 ± 10.22	12.93 ± 6.47	16.16 ± 10.22	9.70 ± 7.92
MUV_644	25.86 ± 12.93	22.63 ± 16.48	25.86 ± 12.93	19.39 ± 12.09	6.47 ± 12.93
MUV_652	19.391 ± 6.47	12.93 ± 12.09	19.39 ± 6.47	16.16 ± 10.22	6.47 ± 7.92

<b>MUV_689</b>	19.39 ± 12.09	12.93 ± 12.09	19.39 ± 12.09	12.93 ± 6.47	6.47 ± 7.92
<b>MUV_692</b>	0	0	0	0	3.23 ± 6.46
<b>MUV_712</b>	25.85 ± 16.48	16.16 ± 14.45	25.85 ± 16.48	22.62 ± 28.17	0
<b>MUV_713</b>	22.63 ± 7.92	16.16 ± 0	22.63 ± 7.92	16.16 ± 10.22	0
<b>MUV_733</b>	9.70 ± 12.93	9.70 ± 7.92	9.70 ± 12.93	12.93 ± 12.09	0
<b>MUV_737</b>	9.70 ± 7.92	16.16 ± 14.46	12.93 ± 6.47	12.93 ± 12.09	0
<b>MUV_810</b>	3.23 ± 6.47	16.16 ± 17.70	3.23 ± 6.47	19.39 ± 15.84	6.47 ± 7.92
<b>MUV_832</b>	45.25 ± 12.09	58.18 ± 7.91	45.25 ± 12.09	61.41 ± 6.46	19.39 ± 12.09
<b>MUV_846</b>	51.72 ± 15.84	54.95 ± 7.92	51.72 ± 15.84	54.95 ± 7.92	9.70 ± 7.92
<b>MUV_852</b>	32.32 ± 10.22	45.25 ± 18.85	32.32 ± 10.22	45.25 ± 18.85	3.23 ± 6.47
<b>MUV_858</b>	9.70 ± 12.93	12.93 ± 12.09	9.70 ± 12.93	16.16 ± 10.22	0
<b>MUV_859</b>	0	6.47 ± 7.92	0	9.70 ± 7.92	0

The results indicate several changes compared to the earlier data in Round I. Significant improvements in EF values can be observed for certain targets.

For example, MUV\_548 showed a notable increase in EF value, rising from 16.16 to 35.56 in MD alone, and from 19.39 to 38.79 for the combined MD+4DES model, highlighting a substantial enhancement in the model's ability to identify active compounds in the top-ranked predictions. Similarly, MUV\_832 exhibited an increase from 45.25 to 58.18 and 61.41 ± 6.46, for MD alone and MD+4DES, respectively, indicating improved performance during the second round of experiments. Overall, the combined MD+4DES model continues to outperform individual descriptor models across multiple targets, though some exceptions indicate variability in its effectiveness depending on the specific target.

From AUC and 1%EF analyses, we can observe overall better performance in Round II, with improvements achieved for 14 out of 17 targets in the MD method (excluding MUV\_652, MUV\_852, and MUV\_858) and for 15 out of 17 targets in the MD+4DES method (excluding MUV\_652 and MUV\_852).

After these analyses, we further employed the SHAP method to identify the 25 most significant descriptors for each target. This additional dimensionality reduction was pursued for two reasons: first, to identify more redundant features, making the representation as compact as possible, and second, to improve the performance by increasing ROC-AUC and EF

values. A smaller descriptor set is particularly useful when working with very large virtual screening datasets.

This process resulted in new matrices for each target, each consisting of 25 MD. We then utilized these matrices to rerun our model, evaluating it using a 3-fold cross-validation approach. This evaluation was conducted by splitting the MUV targets in two different ways: by numeric order and by structural identity. The lists of the top 25 descriptors for each target, derived from both the MD alone and the combined MD+4DES approaches, are reported in Supplementary Material S1, Table S11.

#### 6.4.5. Top 25 Molecular Descriptors

The MD that appeared most frequently in the top 25 across various targets are listed in Table 23. These descriptors play a crucial role in characterizing the physicochemical properties of the molecules and in enhancing the predictive capabilities of the model. Among the most significant descriptors are those related to the Spatial, Electrostatic, Spectral, and Topological properties of the molecules, all of which greatly contribute to classifying molecules as either active or inactive.

**Table 23.** List of most frequent MD the top 25 across various targets, categorized by their type (topological, electrostatic, spatial, and spectral).

MD Class	MD code	MD name
TOPOLOGICAL DESCRIPTOR	58000000	Average Information content (order 1)
TOPOLOGICAL DESCRIPTOR	128000000	nhigh lowest atom weighted BCUTS
TOPOLOGICAL DESCRIPTOR	130000000	nhigh lowest partial charge weighted BCUTS
TOPOLOGICAL DESCRIPTOR	137000000	CarbonType-C1SP2
TOPOLOGICAL DESCRIPTOR	139000000	CarbonType-C3SP2
ELECTROSTATIC DESCRIPTOR	167000000	RNCS Relative negative charged SA (SAMNEG*RNCG) (Zefirov PC)
ELECTROSTATIC DESCRIPTOR	174000000	Min partial charge (Zefirov) for atoms for atom H
ELECTROSTATIC DESCRIPTOR	177000000	Min partial charge (Zefirov) for all atom types
ELECTROSTATIC DESCRIPTOR	605000000	eccentricConnectivityIndex

TOPOLOGICAL DESCRIPTOR	1365000000	[CD2H](=*)-*
TOPOLOGICAL DESCRIPTOR	1368000000	[CD3H0](=*)(-*)-*
TOPOLOGICAL DESCRIPTOR	633000000	[ND2H](-*)-*
TOPOLOGICAL DESCRIPTOR	1372000000	[ND2H0](=*)-*
TOPOLOGICAL DESCRIPTOR	639000000	[ND3H0](-*)(-*)-*
TOPOLOGICAL DESCRIPTOR	641000000	[N,nD3H0](:*)(:*)-,:*
TOPOLOGICAL DESCRIPTOR	645000000	[OD2H0](-*)-*
TOPOLOGICAL DESCRIPTOR	646000000	[O,oD2H0](:*):*
TOPOLOGICAL DESCRIPTOR	1375000000	[SD1H0]=*
TOPOLOGICAL DESCRIPTOR	690000000	LargestPISystem
TOPOLOGICAL DESCRIPTOR	693000000	longestAliphaticChain
TOPOLOGICAL DESCRIPTOR	716000000	PetitJeanShapeGeometrical
TOPOLOGICAL DESCRIPTOR	723000000	NAromaticRings
SPATIAL DESCRIPTOR	884000000	Average Broto-Moreau autocorrelation lag 4 weighted by mass
SPATIAL DESCRIPTOR	902000000	Average Broto-Moreau autocorrelation lag 4 weighted by Sanderson electronegativities
SPATIAL DESCRIPTOR	928000000	Average Broto-Moreau autocorrelation lag 3 weighted by I-state
SPATIAL DESCRIPTOR	945000000	Centered Broto-Moreau autocorrelation lag 2 weighted by mass
SPATIAL DESCRIPTOR	987000000	Centered Broto-Moreau autocorrelation lag 8 weighted by first ionization potential
SPATIAL DESCRIPTOR	1040000000	Average centered Broto-Moreau autocorrelation lag 7 weighted by polarizabilities
SPATIAL DESCRIPTOR	1113000000	Moran autocorrelation lag 6 weighted by I-state
SPATIAL DESCRIPTOR	1119000000	Geary autocorrelation lag 4 weighted by charges
SPATIAL DESCRIPTOR	1165000000	Geary autocorrelation lag 2 weighted by I-state



SPECTRAL DESCRIPTOR	1182000000	Smallest absolute eigenvalue of Burden modified matrix - n 3 weighted by relative mass
SPECTRAL DESCRIPTOR	1185000000	Smallest absolute eigenvalue of Burden modified matrix - n 6 weighted by relative mass
SPECTRAL DESCRIPTOR	1234000000	Smallest absolute eigenvalue of Burden modified matrix - n 7 weighted by relative polarizabilities
SPECTRAL DESCRIPTOR	1253000000	Largest absolute eigenvalue of Burden modified matrix - n 2 weighted by relative I-state
SPECTRAL DESCRIPTOR	1256000000	Largest absolute eigenvalue of Burden modified matrix - n 5 weighted by relative I-state
SPECTRAL DESCRIPTOR	1261000000	Smallest absolute eigenvalue of Burden modified matrix - n 2 weighted by relative I-state

Topological Descriptors focus on the molecular structure and capture connectivity within the molecules without considering the spatial arrangement. The Average Information Content (order 1) evaluates atomic diversity based on connectivity, reflecting molecular complexity. BCUT descriptors, such as `nhigh` lowest atom weighted and `nhigh` lowest partial charge weighted, highlight key features like atomic and charge distributions. `CarbonType` descriptors classify carbon atoms based on hybridization, while specific substructures indicated as SMART represent distinct atomic groups. Other relevant descriptors, such as `LargestPISystem` and `NAromaticRings`, focus on the presence and extent of conjugated or aromatic systems, whereas `longestAliphaticChain` measures the length of unbranched carbon chains. Additionally, `PetitJeanShapeGeometrical` evaluates molecular shape and symmetry, providing essential insights into molecular architecture.

Spatial Descriptors, including Broto-Moreau, Moran, and Geary descriptors, evaluate how specific molecular properties (e.g., mass, charge, electronegativity) are distributed and correlated throughout the molecule. Broto-Moreau Autocorrelation Descriptors measure the correlation of molecular properties (such as mass, electronegativity, polarizability, ionization potential) between atoms separated by a specific number of bonds (lag). Depending on the weighting property, they can capture how these properties vary locally (short lag) or over longer molecular distances (larger lag), providing insights into structural and electronic effects within the molecule.

Moran Autocorrelation Descriptors are statistical measures that quantify the autocorrelation of molecular properties at different points within the molecular graph. Weighted by

properties like I-state (intrinsic state), they reflect how a specific property is distributed throughout the molecule, focusing on whether similar values are clustered or spread apart.

Geary Autocorrelation Descriptors also spatial correlation measures but focus more on the dissimilarity between neighboring atoms' properties. Like Moran descriptors, they are weighted by specific molecular properties (e.g., charges or I-state) but emphasize local variations and differences between atoms, offering a complementary view to Moran's more global correlation.

Spectral descriptors are derived from the eigenvalues of the Burden modified matrix, which represents the electronic and structural properties of a molecule. The matrix is constructed using atomic properties (such as mass, polarizability, or intrinsic valence state) and bond orders between atoms. These descriptors capture important molecular characteristics by analyzing both the smallest and largest eigenvalues of this matrix.

The Smallest Absolute Eigenvalue of Burden Modified Matrix descriptors reflect the lower-energy electronic interactions within the molecule, while the Largest Absolute Eigenvalue of Burden Modified Matrix descriptors represent the higher-energy interactions within the molecule.

Electrostatic descriptors also play a crucial role in understanding molecular interactions. For example, the electrostatic descriptor Relative Negative Charged Surface Area quantifies the surface area of negatively charged atoms, influencing chemical reactivity and physical properties. The minimum partial charges for hydrogen and all atom types, as defined by Zefirov, provide insight into charge distribution, and help identify reactive sites within the molecule.

#### **6.4.6. 3-Fold Cross-Validation: Evaluating Model Performance with Top Descriptors**

In order to thoroughly evaluate the performance of the model built using the top 25 molecular descriptors, we employed two distinct methods for splitting the MUV targets during the 3-fold cross-validation process. This approach allows for a more comprehensive assessment of the models' predictive capabilities under varying data distributions. The first method utilized numeric order (according to the target's MUV ID) for target splitting, while the second method was based on structural identity (according to FASTA sequences). Each method provides

unique insights into how the selected descriptors influence model performance, which will be discussed in the following subsections. The aim of these splitting strategies is to understand whether reducing the features and utilizing those involved in a specific subset of targets (training set) can affect performance when the model is subsequently evaluated on a different subset of targets (test set).

### 6.4.6.1. Insights from 3-Fold Cross-Validation with Numeric Order Splitting

To evaluate the model performance using the numeric order splitting method, the MUV targets were divided into three distinct groups based on their MUV IDs. The first group comprised of 6 targets: MUV\_466, MUV\_548, MUV\_600, MUV\_644, MUV\_652, and MUV\_689 (indicated in yellow in Figure 68). The second group included 6 targets which were MUV\_692, MUV\_712, MUV\_713, MUV\_733, MUV\_737, and MUV\_810 (indicated in green in Figure 68). Lastly, the third group consisted of the remaining 5 MUV targets: MUV\_832, MUV\_846, MUV\_852, MUV\_858, and MUV\_859 (indicated in red in Figure 68). This partition allowed us to create different scenarios for training and test sets for the 3-fold cross-validation process.

The training and testing scenarios were as follows:

1. **Training Set:** Group 1 + Group 2 - **Test Set:** Group 3
2. **Training Set:** Group 1 + Group 3 - **Test Set:** Group 2
3. **Training Set:** Group 2 + Group 3 - **Test Set:** Group 1

#### MUV: Numeric Order Splitting

MUV ID	Target	Interaction	MUV ID	Target	Interaction	MUV ID	Target	Interaction
466	S1P1 receptor	agonists	692	SF1	agonists	832	Cathepsin G	inhibitors
548	PKA	inhibitors	712	HSP 90	inhibitors	846	FXIa	inhibitors
600	SF1	inhibitors	713	ER- $\alpha$ -coact. bind.	inhibitors	852	FXIIa	inhibitors
644	Rho-Kinase2	inhibitors	733	ER- $\beta$ -coact. bind.	inhibitors	858	D1 receptor	Allosteric modulators
652	HIV RT-RNase	inhibitors	737	ER- $\alpha$ -coact. bind.	potentiators	859	M1 receptor	Allosteric inhibitors
689	Eph rec. A4	inhibitors	810	FAK	inhibitors			

<b>TRAINING SET 1</b>	1 + 2	<b>TEST SET</b>	3
<b>TRAINING SET 2</b>	1 + 3	<b>TEST SET</b>	2
<b>TRAINING SET 3</b>	2 + 3	<b>TEST SET</b>	1

**Figure 68.** Overview of the MUV target division into three groups based on MUV IDs and the corresponding combinations used for training and test sets during the 3-fold cross-validation. Group 1 is highlighted in yellow, while groups 2 and 3 in green and red, respectively.

For each of the 17 MUV targets, we generated two distinct lists of 25 descriptors using the SHAP method based on the top 50 descriptors from the previous step. The first list was derived solely from MD, capturing the most relevant features identified through the MD-based analysis. The second list was obtained by combining MD with the 15 4DES descriptors, leading to a refined selection of features that integrate both MD and 4DES information (lists reported in Supplementary Material S1, Table S11). We computed the Reciprocal Rank (RR) for both descriptor lists across each training set. RR is a metric used to evaluate the position of an item within a ranked list. It is calculated as the inverse of the item's rank position. The idea is to give more weight to descriptors that appear higher (i.e., have lower rank numbers) in the list. This is particularly useful when aggregating rankings from multiple lists and identify descriptors that consistently perform well across different models or methods.

This calculation not only allowed us to assess the consistency and rank correlation between the descriptors identified by the MD and MD+4DES methods, but also helped to identify which descriptors appeared most frequently across the lists. This led to two diverse ranking for each training set (Calculations are available at Supplementary Material S2).

With these descriptor lists in hand, we proceeded to evaluate the model performance on the TEST sets across different experimental scenarios. In the first set of experiments, EXPERIMENT 1 (MD list), we calculated the AUC values for the TEST sets by using the descriptors from the MD list. For EXPERIMENT 2 (MD list  $\cup$  MD+4DES list), we employed a different strategy by utilizing all the descriptors from both lists, ensuring that no descriptor was counted more than once. This meant that we included all unique descriptors as well as those that were shared between the MD and MD+4DES lists, but we avoided duplicating any descriptor that appeared in both lists. This approach allowed us to assess the overall model performance on the test sets based on a comprehensive set of descriptors, incorporating both common and unique features without redundancy. In EXPERIMENT 3 (MD+4DES list) we utilized descriptors from the MD+4DES list, while for EXPERIMENT 4 (MD list  $\cap$  MD+4DES list), just descriptors in common between MD list and MD+4DES list

In each experiment we used the descriptors in two ways, once alone e once combined with 4DES. Results for each Experiment are provided in Supplementary Material S3

### 6.4.6.2. Insights from 3-Fold Cross-Validation with Structural Identity Splitting

To evaluate the model performance using the structural identity splitting method, the MUV targets were divided into three distinct groups based on their FASTA sequence.

This process was carried out in multiple steps. **Step 1** involved generating a linkage matrix using the “average” method, also known as UPGMA (Unweighted Pair Group Method with Arithmetic mean). The linkage matrix was constructed by calculating pairwise distances between all targets based on their FASTA sequences, which were then averaged to determine the distances between clusters. This step is critical in hierarchical clustering, as it provides a numerical representation of the identity between sequences, which is used to progressively merge clusters of targets. Percentage identity values, calculated using MOE (Molecular Operating Environment), are presented in the identity matrix shown in the Figure 69. Each value represents the degree of identity between two targets, with higher values (ranging up to 100) indicating a stronger structural identity, while lower values indicate more distinct sequences.

Target	466	548	600	644	652	689	692	712	713	733	737	810	832	846	852	858	859	C.01	Cluster
600	10.8	8.9	100.0	15.4	11.3	11.9	100.0	13.0	21.5	19.1	21.5	14.3	7.6	12.4	14.3	9.3	13.4	1	1
692	10.8	8.9	100.0	15.4	11.3	11.9	100.0	13.0	21.5	19.1	21.5	14.3	7.6	12.4	14.3	9.3	13.4	1	1
644	3.5	8.4	5.1	100.0	5.0	8.6	5.1	8.1	5.4	4.4	5.4	11.4	2.3	6.3	5.9	4.0	3.8	2	2
548	12.5	100.0	11.7	33.0	14.2	20.2	11.7	14.0	14.5	12.8	14.5	21.4	8.0	11.7	11.4	14.0	10.8	3	2
810	5.9	7.1	6.4	15.1	6.4	10.8	6.4	10.2	7.6	7.0	7.6	100.0	2.8	8.6	8.2	6.1	6.0	4	2
689	5.2	6.9	5.6	11.9	6.7	100.0	5.6	9.5	7.1	7.9	7.1	11.7	3.0	8.1	7.9	6.3	6.0	5	2
858	14.3	11.0	10.3	12.1	14.1	13.0	10.3	13.0	9.4	10.3	9.4	14.3	8.1	9.2	10.8	100.0	22.2	6	3
859	13.3	8.3	12.8	11.5	13.0	12.8	12.8	14.6	13.0	11.1	13.0	13.7	7.0	12.6	13.7	21.5	100.0	7	3
466	100.0	11.5	13.1	12.8	13.1	15.2	13.1	12.6	12.6	12.0	12.6	16.2	7.1	14.4	13.1	16.5	16.0	8	3
652	8.9	8.9	9.3	12.5	100.0	11.8	9.3	13.4	12.0	10.9	12.0	12.0	6.1	13.8	11.8	11.3	10.7	9	3
712	6.6	6.7	8.2	15.3	10.0	12.8	8.2	100.0	9.3	9.8	9.3	14.6	3.8	9.0	8.3	7.9	9.0	10	3
846	8.3	6.6	9.1	13.6	12.3	13.0	9.1	10.6	9.6	9.1	9.6	14.4	12.0	100.0	21.4	6.6	8.8	11	4
852	8.1	6.5	10.4	13.8	11.1	12.5	10.4	9.9	11.2	11.4	11.2	14.0	12.4	21.8	100.0	7.8	10.2	12	4
832	10.6	10.6	14.1	12.5	13.3	11.8	14.1	11.0	14.5	14.9	14.5	11.4	100.0	29.4	29.8	14.1	12.5	13	4
713	8.1	8.6	16.5	12.4	11.4	12.1	16.5	11.4	100.0	44.4	100.0	13.4	6.2	10.1	12.1	7.1	9.9	14	5
737	8.1	8.6	16.5	12.4	11.4	12.1	16.5	11.4	100.0	44.4	100.0	13.4	6.2	10.1	12.1	7.1	9.9	14	5
733	8.7	8.5	16.6	11.9	11.5	14.7	16.6	13.8	50.0	100.0	50.0	14.0	7.2	10.8	12.8	8.7	9.2	15	5

Figure 69. Identity matrix illustrating the pairwise similarity values between MUV targets based on their FASTA sequences.

**Step 2** involved forming flat clusters starting from the linkage matrix.

Using the Scipy hierarchical clustering function known as “fcluster function” with the criterion set to 'distance' and a threshold value of 0.1, the dendrogram generated in Step 1 was cut to

create distinct clusters. The distance criterion  $d=0.1$  was chosen to ensure that clusters formed include targets that have an identity level above this threshold, effectively grouping highly similar targets together. This allows for the formation of three main clusters, where each cluster contains targets that are similar based on this distance cutoff.

**Step 3** involved forming flat clusters from the linkage matrix with the goal of limiting the total number of clusters. In this step, the `fcluster` function was used again, but with the criterion set to “maxclust” and the threshold value set to 5. This ensured that no more than five clusters were formed. The aim was to create a broader categorization of the targets, providing an additional layer of grouping that can be useful for comparing model performance across different levels of clustering resolution.

These two methods led to divide the 17 targets into three main clusters, with the goal of evaluating how the model performs across different levels of structural identity. This structural identity splitting allows for an unbiased validation process, ensuring that targets with high similarity are grouped together, which facilitates a more rigorous assessment of the model's predictive capabilities across structurally similar and dissimilar targets. Essentially, this step can be seen as a generalizability assessment to see how MD perform when reducing the similarity between training and testing compounds.

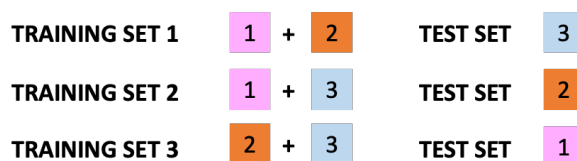
The first group comprised of targets MUV\_548, MUV\_600, MUV\_644, MUV\_689, MUV\_692, and MUV\_810 (indicated in purple in Figure 70). The second group included targets MUV\_466, MUV\_652, MUV\_712, MUV\_858, and MUV\_859 (indicated in orange in Figure 70). Lastly, the third group consisted of targets MUV\_713, MUV\_733, MUV\_737, MUV\_832, MUV\_846 and MUV\_852 (indicated in blue in Figure 70). This partition allowed us to create different protocols of training and test sets for the 3-fold cross-validation process.

The combinations were as follows:

1. **Training Set:** Group 1 + Group 2 - **Test Set:** Group 3
2. **Training Set:** Group 1 + Group 3 - **Test Set:** Group 2
3. **Training Set:** Group 2 + Group 3 - **Test Set:** Group 1

## MUV: Structural Identity Splitting

MUV ID	Target	Interaction	MUV ID	Target	Interaction	MUV ID	Target	Interaction
548	PKA	inhibitors	466	S1P1 receptor	agonists	713	ER- $\alpha$ -coact. bind.	inhibitors
600	SF1	inhibitors	652	HIV RT-RNase	inhibitors	733	ER- $\beta$ -coact. bind.	inhibitors
644	Rho-Kinase2	inhibitors	712	HSP 90	inhibitors	737	ER- $\alpha$ -coact. bind.	potentiators
689	Eph rec. A4	inhibitors	858	D1 receptor	Allosteric modulators	832	Cathepsin G	inhibitors
692	SF1	agonists	859	M1 receptor	Allosteric inhibitors	846	FXIa	inhibitors
810	FAK	inhibitors				852	FXIIa	inhibitors



**Figure 70.** Overview of the MUV target division into three groups based on FASTA sequences and the corresponding combinations used for training and test sets during the 3-fold cross-validation. Group 1 is highlighted in purple, while groups 2 and 3 in orange and blue, respectively.

As described in section 6.4.5.1., we computed the reciprocal rank for both descriptor lists (Supplementary Material S1, Table S11) across each training set, resulting in six rankings (two per target), as detailed in Supplementary Material S4. Using these descriptor lists, we evaluated the model performance on the TEST sets under identical experimental scenarios described earlier. In each experiment, descriptors were applied both independently and in combination with 4DES, allowing for a comprehensive performance assessment across different experimental protocols. Detailed results are provided in Supplementary Material S5, and the next section contains a summary of the important results.

### 6.4.7. Analysis of results across different Experimental scenarios

To analyze the results of the experiments conducted using the two splitting methods (numeric order splitting and structural identity splitting), we can compare the average model performance in terms of AUC across the different experimental scenarios, which is depicted in Table 24.

**Table 24.** Average AUC values for model performance across different experimental scenarios, comparing numeric order splitting and structural identity-based splitting methods. Results are shown for each experiment, including the use of different descriptor lists (MD, MD+4DES, their union, and their intersection).

<b>EXPERIMENTS</b>	<b>Structural Identity Splitting</b>	<b>Numeric Order Splitting</b>
<b>Experiment 1</b>	<b>MEAN VALUES</b>	<b>MEAN VALUES</b>
(MD list)	0.703 ± 0.130	0.695 ± 0.108
MD list + 4DES	0.715 ± 0.125	0.707 ± 0.113
<b>Experiment 2</b>	<b>MEAN VALUES</b>	<b>MEAN VALUES</b>
(MD list ∪ MD+4DES list)	0.750 ± 0.112	0.724 ± 0.112
(MD list ∪ MD+4DES list) + 4DES	0.750 ± 0.120	0.730 ± 0.107
<b>Experiment 3</b>	<b>MEAN VALUES</b>	<b>MEAN VALUES</b>
(MD+4DES list)	0.742 ± 0.118	0.695 ± 0.147
(MD+4DES list) + 4DES	0.749 ± 0.116	0.725 ± 0.110
<b>Experiment 4</b>	<b>MEAN VALUES</b>	<b>MEAN VALUES</b>
(MD list ∩ MD+4DES list)	0.664 ± 0.123	0.652 ± 0.123
(MD list ∩ MD+4DES list) + 4DES	0.723 ± 0.100	0.697 ± 0.114

Observing the mean values obtained, we note that the structural identity-based splitting tends to yield slightly better results compared to numeric order splitting in all experiments. This suggests that splitting based on structural identity is more suitable for capturing the relationships between targets, thereby improving the predictive ability of the model.

Experiment 1 (MD list) shows very similar results between the two approaches, with an average AUC value of  $0.703 \pm 0.130$  for structural identity and  $0.695 \pm 0.108$  for numeric order. This indicates that using MD descriptors alone is not significantly affected by the splitting method. However, the inclusion of 4DES descriptors (MD list + 4DES) slightly improves the performance for both approaches, with a greater advantage seen in the structural identity-based splitting. In Experiment 2, where all descriptors from the MD and MD+4DES lists were combined, we observe that the structural identity-based approach achieved a mean value of  $0.750 \pm 0.112$ , higher than  $0.724 \pm 0.112$  obtained with numeric order splitting. Including the 4DES descriptors further improved performance in both cases, suggesting that including 4D shape based molecular descriptors lead to better model discriminative ability between targets. In Experiment 3, where descriptors from the MD+4DES



list were used, structural identity-based splitting continued to show an advantage over numeric order splitting, with a mean value of  $0.742 \pm 0.118$  compared to  $0.695 \pm 0.147$ . Again, using 4DES descriptors improved performance, resulting in  $0.749 \pm 0.116$  for structural identity versus  $0.725 \pm 0.110$  for numeric order. Finally, Experiment 4, which uses the intersection of the two descriptor lists, showed a mean value of  $0.664 \pm 0.123$  versus  $0.652 \pm 0.123$ . Including the 4DES descriptors improved both approaches.

The structural identity-based splitting introduces greater complexity to the model, as it involves training on certain groups of structures and testing on entirely different ones. This requires the model to generalize across different structural identities, which is a more challenging scenario compared to numeric order splitting. Despite this added complexity, the structural identity-based approach consistently outperforms the numeric order splitting, demonstrating superior predictive performance in all experimental scenarios and indicating the power of our descriptors when combined with SVM model.

Additionally, among all experiments, Experiment 2 consistently achieves the best results, indicating that the combination of all descriptors from the MD and MD+4DES lists provides the most comprehensive and informative feature set for model prediction.

#### **6.4.8. Model Performance examination on LIT-PCBA dataset**

To further assess the robustness and generalizability of the model developed using the top molecular descriptors, we conducted an evaluation on the LIT-PCBA dataset. This dataset comprises a diverse array of biological targets and serves as an excellent benchmark for evaluating predictive models in the context of drug discovery and molecular interactions as it contains less bias in comparison with MUV and other datasets like DUD-E. We employed the same experimental framework, as outlined in previous sections, and tested the model on descriptors used during Experiment 2 on a subset of LIT-PCBA. In details we decided to utilize a target from LIT-PCBA (MAPK1) for which experimental and comparable data were already available in literature, and other two targets (PPARG and TP53) for evaluating performance. The number of active and inactive compounds within Validation and Training sets for each MAPK1, PPARG, and TP53 are provided in Table 25.

**Table 25.** LIT-PCBA Subset of 3 targets utilized in the final evaluation of the model.

TARGET ID	TARGET NAME	ACTIVES		INACTIVES	
		Validation set	Training set	Validation set	Training set
PPARG	Peroxisome proliferator-activated receptor $\gamma$	6	21	1,302	3,909
TP53	Cellular tumor antigen p53	19	60	1,042	3,126
MAPK1	Mitogen-activated protein kinase 1	77	231	15,657	46,972

Unlike previous experiments where a 5-fold cross-validation was employed, the evaluation on the LIT-PCBA dataset did not require such an approach. This is because the LIT-PCBA dataset already provides a predefined split into training and validation sets, which we used for the model assessment. This modification to our script allowed us to directly use the provided partitioning, ensuring consistency with the standard practices for this dataset.

AUC and 1%EF values for the LIT-PCBA subset, when using MD alone, 4DES alone and the combined MD+4DES approach, are shown in Table 26.

**Table 26.** Performance of the model on the LIT-PCBA dataset for three targets (MAPK1, PPARG, TP53), showing AUC and 1%EF values for three descriptor approaches: MD, 4DES and MD+4DES combined.

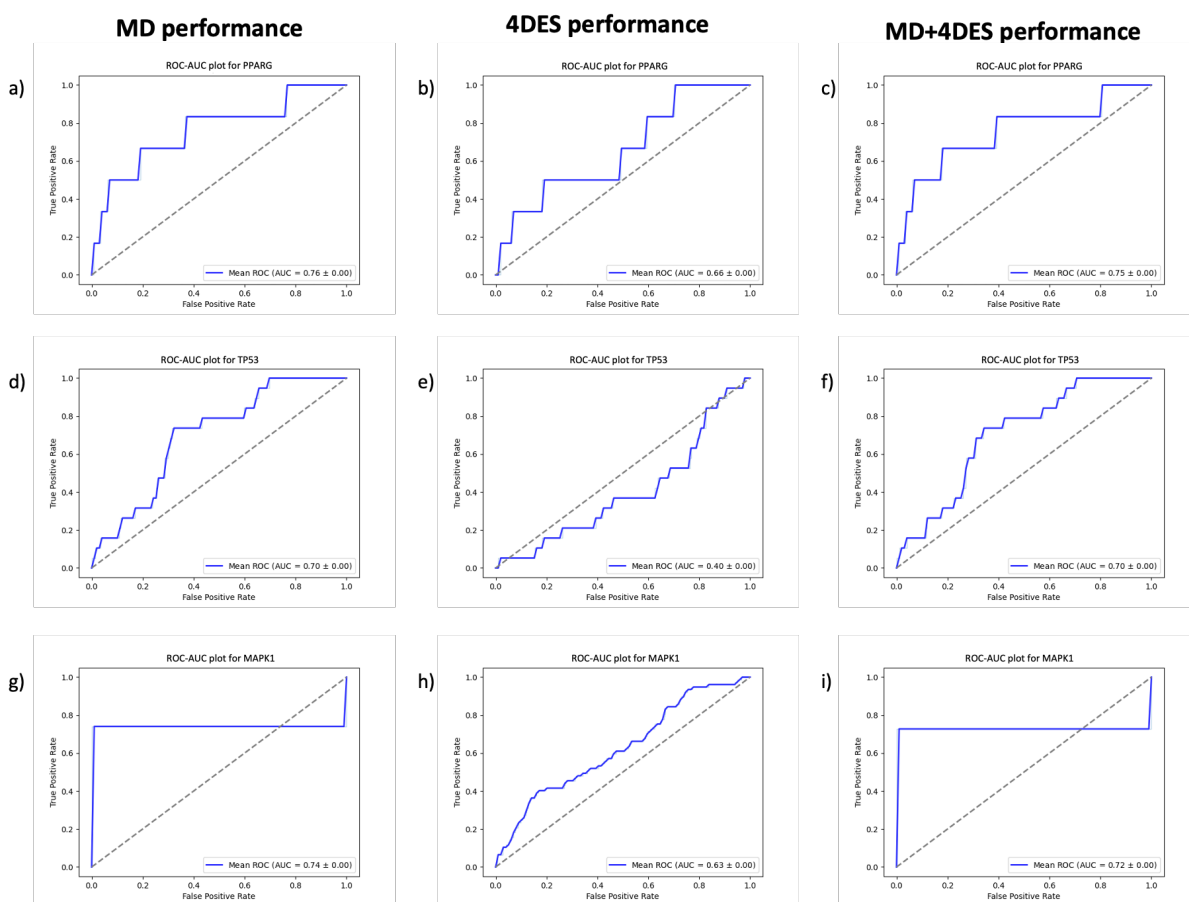
TARGET ID	MD		4DES		MD+4DES	
	AUC	1%EF	AUC	1%EF	AUC	1%EF
PPARG	0.762	16.654	0.658	0	0.754	16.654
TP53	0.700	5.072	0.402	0	0.700	5.072
MAPK1	0.740	73.595	0.626	6.456	0.727	72.304

For MAPK1, the model achieved an AUC of 0.740 with the MD approach, 0.626 with 4DES, and 0.727 with the combined MD+4DES method. These results demonstrate the model's robustness, particularly with the MD descriptor, which yielded the best predictive performance. Additionally, the model showed impressive values for 1%EF with 73.595 for MD and 72.304 for MD+4DES. This indicates a strong ability to enrich active compounds early in the ranking process, which is crucial in virtual screening. These results align well with the

literature, where top-performing models like EGT+TGT-At-DP [255] shows an AUC comparable of 0.743 (similar to ours), while other models, such as GLAM [256], TransformerCPI [257], and DGraphDTA [258] are underperforming with AUC values of 0.730, 0.683, and 0.665, respectively. It should be noted that our approach only utilized SVM with default parameters, unlike the results reported in [256-258] where they used sophisticated deep learning methods. Hence, we hypothesize that a more sophisticated ML classifier would help in boosting the results reported here.

For PPARG, the model achieved an AUC of 0.762 with MD, with a high 1%EF of 16.654. This is a strong result, indicating that the model successfully enriched the most active compounds early in the ranking. The combined MD+4DES approach produced a very similar AUC (0.754) and the same 1%EF value, implying that both methods were effective. These values show a solid ability to identify active compounds for PPARG, though the performance could still be benchmarked against the top models in literature.

For TP53, the model achieved an AUC of 0.700 with MD, which is fairly strong, and a moderate 1%EF of 5.072 for both MD alone and MD+4DES combined. ROC-AUC for PPARG, TP53 and MAPK1 are shown in Figure 71.



**Figure 71.** (a) ROC-AUC for target PPARG when using the MD approach; (b) ROC-AUC for target PPARG when using the 4DES approach; (c) ROC-AUC for target PPARG when using the MD+4DES combined approach; (d) ROC-AUC for target TP53 when using the MD approach; (e) ROC-AUC for target TP53 when using 4DES approach; (f) ROC-AUC for target TP53 when the using the MD+4DES combined approach; (g) ROC-AUC for target MAPK1 when using the MD approach; (h) ROC-AUC for target MAPK1 when using the 4DES approach; (i) ROC-AUC target MAPK1 when using the MD+4DES combined approach.

In summary, the model demonstrated a good performance for MAPK1 and PPARG, yielding competitive, if not better than literature, AUC and 1%EF values. This suggests a high capacity for accurate early enrichment of active compounds, which is crucial in virtual screening efforts. TP53, on the other hand, showed a solid yet less impressive performance compared to the other two targets. This could be attributed to the inherent complexity of this protein, which is known for its challenging ligand-binding properties.

The MD approach consistently performed the best, particularly for MAPK1, where it achieved the highest AUC and 1%EF scores, closely aligned with top-performing models from the literature. The MD+4DES combination offered slight improvements for some targets but was generally not as robust as MD. 4DES alone performs poorly when used for LIT-PCBA,

reinforcing the superiority of molecular descriptors (MD) in capturing relevant features for predicting active compounds across these targets.

## 6.5. Conclusions and Future Perspectives

ML has become a powerful tool in drug discovery, enabling the analysis of large datasets to predict molecular activity and identify promising therapeutic candidates. A crucial aspect of applying ML in this field is the accurate representation of molecular structures, as this directly impacts the model's predictive performance.

Results reported in this section has demonstrated the effectiveness of combining MD and 4DES for predicting molecular activity against specific biological targets. The results show that integrating molecular descriptors with electrostatic and spatial descriptors improves the predictive capabilities of ML models, increasing both the AUC and the EF compared to using individual methods. The performance analysis of the proposed approach, through a series of experiments on benchmark datasets (such as MUV and LIT-PCBA), highlighted the importance of feature selection. Indeed, the most significant finding is the impact of feature reduction on model performance. The use of SHAP significantly enhanced model AUC and 1%EF values, confirming that eliminating redundant or non-informative descriptors can increase the model's discriminative power. By reducing the feature set to only the most relevant descriptors, the model not only became more efficient but also exhibited marked improvements in predictive accuracy.

The analysis conducted on different three-fold cross-validation scenarios (based on numerical order and structural identity) suggest that the structure-based splitting approach is more effective in achieving high generalizability of our proposed ML approach. Finally, the application of the model on the LIT-PCBA dataset showed promising results, highlighting the effectiveness of the proposed framework for advancing virtual screening in drug discovery.

There are several promising research avenues for future exploration. In this vein, future work includes experimenting with additional targets from the LIT-PCBA dataset, as well as other datasets, to validate and expand the applicability of the proposed MD descriptor and ML model. Moreover, other ML frameworks beyond SVM could be developed to assess their performance against the current model. Furthermore, there is potential for identifying new molecular descriptors that could be integrated with the existing ES descriptors, enriching the feature space, and potentially leading to even more robust predictive models.

## 6.6. Materials and Methods

### 6.6.1. Benchmark Datasets

In this study, two benchmark datasets were used: the MUV dataset [221] and the LIT-PCBA dataset [222]. The MUV dataset includes assay data for 17 biological targets designed to minimize analogue bias and artificial enrichment, thus providing a reliable foundation for training, and validating the ML models. The LIT-PCBA dataset was used for evaluating the generalizability of the model, as it presents challenges related to diverse biological targets and minimized potency distribution biases among active compounds.

### 6.6.2. Molecular Representation

Two sets of descriptors were employed to represent the molecular structures in the ML models: MD and 4DES. MD were calculated using the MOLDESTO software [108], capturing approximately 1,000 physicochemical properties, such as topological, geometric, electronic, and thermodynamic features. 4DES, on the other hand, were generated using EShape [220]. These descriptors provide complementary data that help improve the classification of active and inactive compounds.

### 6.6.3. Features Selection

To optimize the performance of the ML models, feature selection was carried out using SHAP [254]. This approach identified the most important descriptors for each target, reducing the dimensionality of the feature set and enhancing interpretability. Roughly speaking, SHAP works by calculating the contribution of each feature to a model's prediction using concepts from game theory. It assigns each feature an importance value that represents how much that feature changes the utilized ML model's output when included, averaged over all possible combinations of other features. We direct interested readers to see [254] for a detailed mathematical description on how SHAP works. We used SHAP in two rounds of analysis, selecting 50 and then 25 top most contributing descriptors based on their importance, to refine and improve model's performance.

#### **6.6.4. Machine Learning Model Development**

A SVM model was developed to classify compounds as active or inactive against specific biological targets. The SVM model was chosen for its efficiency in handling high-dimensional datasets and binary classification tasks. The performance of the model was evaluated using 5-fold cross-validation on the MUV dataset. Stratified cross-validation ensured that each fold contained a balanced distribution of active and inactive compounds. The model was further refined using the selected top descriptors, and a 3-fold cross-validation approach was employed to test the robustness and generalizability of the model under different experimental conditions.

#### **6.6.5. Evaluation Metrics**

The ROC curve, the AUC, and the 1%EF were used to evaluate the performance of the model. The AUC provides an indication of the model's overall discriminative ability, while the 1%EF is a critical metric in virtual screening, assessing the model's capability to prioritize active compounds among the top-ranked predictions. The enrichment factor was calculated for the top 1% of predictions, giving insights into the effectiveness of the model in early retrieval of active compounds.

#### **6.6.6. Structural Identity Splitting**

The structural identity splitting was executed in multiple steps to ensure effective clustering of similar sequences, as follows below:

##### **6.6.6.1. Step 1: Linkage Matrix Generation**

The linkage matrix was built by utilizing the "average" method, also known as the UPGMA. This involved calculating pairwise distances between all MUV targets based on their FASTA sequences. The average distance between clusters was determined, providing a numerical representation of the identity between sequences. Higher percentage identity values, calculated using MOE, indicate greater structural similarity, while lower values reflect more distinct sequences.



### **6.6.6.2. Step 2: Forming Flat Clusters**

Next, the `fcluster` function, from the SciPy library in python, was used to form flat clusters based on the linkage matrix generated in Step 1. The criterion was set to “distance”, with a threshold value of 0.1. This threshold was chosen to ensure that only targets with an identity level above this cutoff were grouped together. This method facilitated the formation of three main clusters.

### **6.6.6.3. Step 3: Limiting Cluster Formation**

To further refine the clustering approach, the `fcluster` function was employed again, this time setting the criterion to "maxclust" with a maximum threshold of 5 clusters. This adjustment aimed to limit the total number of clusters formed, providing a broader categorization of the targets while still enabling effective comparisons of model performance across different levels of structural identity.

### **6.6.7. Reciprocal Rank**

For each of the 17 MUV targets, we calculated the RR for two distinct lists of 25 descriptors derived from the SHAP method. The RR for each descriptor was computed as the inverse of its rank position in the list. This approach allowed us to emphasize the importance of higher-ranked descriptors and assess their consistency across different ranking methods. We then aggregated the RR values for each descriptor across both lists to identify those that appeared most frequently at the top of the rankings, facilitating a comprehensive evaluation of their predictive significance.

### **6.6.8. Validation on the LIT-PCBA Dataset**

The model developed using the MUV dataset was validated on the LIT-PCBA dataset to assess its ability to generalize to unseen targets. The LIT-PCBA dataset provided predefined training and validation sets, which allowed for consistent evaluation of the model without the need for cross-validation. Performance was evaluated for specific targets, including MAPK1, PPARG, and TP53, which offered insights into the robustness and effectiveness of the developed model in addressing complex virtual screening challenges.

## 7. Bibliography

1. Wang, L.; Wang, Y.; Ye, D.; Liu, Q. Review of the 2019 novel coronavirus (SARS-CoV-2) based on current evidence. *Int J Antimicrob Agents* **2020**, *55*, 105948, doi:10.1016/j.ijantimicag.2020.105948.
2. Cucinotta, D.; Vanelli, M. WHO Declares COVID-19 a Pandemic. *Acta Biomed* **2020**, *91*, 157-160, doi:10.23750/abm.v91i1.9397.
3. Krishnamoorthy, S.; Swain, B.; Verma, R.S.; Gunthe, S.S. SARS-CoV, MERS-CoV, and 2019-nCoV viruses: an overview of origin, evolution, and genetic variations. *Virusdisease* **2020**, 1-13, doi:10.1007/s13337-020-00632-9.
4. Yu, W.L.; Toh, H.S.; Liao, C.T.; Chang, W.T. Cardiovascular Complications of COVID-19 and Associated Concerns: A Review. *Acta Cardiol Sin* **2021**, *37*, 9-17, doi:10.6515/ACS.202101\_37(1).20200913A.
5. Page, E.M.; Ariëns, R.A.S. Mechanisms of thrombosis and cardiovascular complications in COVID-19. *Thromb Res* **2021**, *200*, 1-8, doi:10.1016/j.thromres.2021.01.005.
6. Liu, L.; Jing, H.; Wu, X.; Xiang, M.; Novakovic, V.A.; Wang, S.; Shi, J. The cross-talk of lung and heart complications in COVID-19: Endothelial cells dysfunction, thrombosis, and treatment. *Front Cardiovasc Med* **2022**, *9*, 957006, doi:10.3389/fcvm.2022.957006.
7. Ghasemiyeh, P.; Mohammadi-Samani, S. Lessons we learned during the past four challenging years in the COVID-19 era: pharmacotherapy, long COVID complications, and vaccine development. *Virology* **2024**, *21*, 98, doi:10.1186/s12985-024-02370-6.
8. Cao, M.; Zhang, D.; Wang, Y.; Lu, Y.; Zhu, X.; Li, Y.; Xue, H.; Lin, Y.; Zhang, M.; Sun, Y.; et al. Clinical Features of Patients Infected with the 2019 Novel Coronavirus (COVID-19) in Shanghai, China. *medRxiv* **2020**, doi:10.1101/2020.03.04.20030395.
9. Martorana, A.; Gentile, C.; Lauria, A. In Silico Insights into the SARS CoV-2 Main Protease Suggest NADH Endogenous Defences in the Control of the Pandemic Coronavirus Infection. *Viruses* **2020**, *12*, doi:10.3390/v12080805.
10. Guan, W.J.; Ni, Z.Y.; Hu, Y.; Liang, W.H.; Ou, C.Q.; He, J.X.; Liu, L.; Shan, H.; Lei, C.L.; Hui, D.S.C.; et al. Clinical Characteristics of Coronavirus Disease 2019 in China. *N Engl J Med* **2020**, *382*, 1708-1720, doi:10.1056/NEJMoa2002032.
11. Lu, X.; Zhang, L.; Du, H.; Zhang, J.; Li, Y.Y.; Qu, J.; Zhang, W.; Wang, Y.; Bao, S.; Li, Y.; et al. SARS-CoV-2 Infection in Children. *N Engl J Med* **2020**, *382*, 1663-1665, doi:10.1056/NEJMc2005073.
12. Bian, L.; Gao, F.; Zhang, J.; He, Q.; Mao, Q.; Xu, M.; Liang, Z. Effects of SARS-CoV-2 variants on vaccine efficacy and response strategies. *Expert Rev Vaccines* **2021**, *20*, 365-373, doi:10.1080/14760584.2021.1903879.
13. Citarella, A.; Scala, A.; Piperno, A.; Micale, N. SARS-CoV-2 M<sup>pro</sup>: A Potential Target for Peptidomimetics and Small-Molecule Inhibitors. *Biomolecules* **2021**, *11*, doi:10.3390/biom11040607.
14. Zhu, H.; Du, W.; Song, M.; Liu, Q.; Herrmann, A.; Huang, Q. Spontaneous binding of potential COVID-19 drugs (Camostat and Nafamostat) to human serine protease TMPRSS2. *Comput Struct Biotechnol J* **2021**, *19*, 467-476, doi:10.1016/j.csbj.2020.12.035.
15. Wang, X.; Cao, R.; Zhang, H.; Liu, J.; Xu, M.; Hu, H.; Li, Y.; Zhao, L.; Li, W.; Sun, X.; et al. The anti-influenza virus drug, arbidol is an efficient inhibitor of SARS-CoV-2 in vitro. *Cell Discov* **2020**, *6*, 28, doi:10.1038/s41421-020-0169-8.

16. Juurlink, D.N. Safety considerations with chloroquine, hydroxychloroquine and azithromycin in the management of SARS-CoV-2 infection. *CMAJ* **2020**, *192*, E450-E453, doi:10.1503/cmaj.200528.
17. Costanzo, M.; De Giglio, M.A.R.; Roviello, G.N. SARS-CoV-2: Recent Reports on Antiviral Therapies Based on Lopinavir/Ritonavir, Darunavir/Umifenovir, Hydroxychloroquine, Remdesivir, Favipiravir and other Drugs for the Treatment of the New Coronavirus. *Curr Med Chem* **2020**, *27*, 4536-4541, doi:10.2174/0929867327666200416131117.
18. Nguyen, H.L.; Thai, N.Q.; Truong, D.T.; Li, M.S. Remdesivir Strongly Binds to Both RNA-Dependent RNA Polymerase and Main Protease of SARS-CoV-2: Evidence from Molecular Simulations. *J Phys Chem B* **2020**, *124*, 11337-11348, doi:10.1021/acs.jpcc.0c07312.
19. Hoffmann, M.; Schroeder, S.; Kleine-Weber, H.; Müller, M.A.; Drosten, C.; Pöhlmann, S. Nafamostat Mesylate Blocks Activation of SARS-CoV-2: New Treatment Option for COVID-19. *Antimicrob Agents Chemother* **2020**, *64*, doi:10.1128/AAC.00754-20.
20. Gunst, J.D.; Staerke, N.B.; Pahus, M.H.; Kristensen, L.H.; Bodilsen, J.; Lohse, N.; Dalgaard, L.S.; Brønnum, D.; Frøbert, O.; Hønge, B.; et al. Efficacy of the TMPRSS2 inhibitor camostat mesilate in patients hospitalized with Covid-19-a double-blind randomized controlled trial. *EClinicalMedicine* **2021**, *35*, 100849, doi:10.1016/j.eclinm.2021.100849.
21. Vankadari, N. Arbidol: A potential antiviral drug for the treatment of SARS-CoV-2 by blocking trimerization of the spike glycoprotein. *Int J Antimicrob Agents* **2020**, *56*, 105998, doi:10.1016/j.ijantimicag.2020.105998.
22. Plaze, M.; Attali, D.; Petit, A.C.; Blatzer, M.; Simon-Lorriere, E.; Vinckier, F.; Cachia, A.; Chrétien, F.; Gaillard, R. Repurposing chlorpromazine to treat COVID-19: The reCoVery study. *Encephale* **2020**, *46*, 169-172, doi:10.1016/j.encep.2020.05.006.
23. Plaze, M.; Attali, D.; Prot, M.; Petit, A.C.; Blatzer, M.; Vinckier, F.; Levillayer, L.; Chiaravalli, J.; Perin-Dureau, F.; Cachia, A.; et al. Inhibition of the replication of SARS-CoV-2 in human cells by the FDA-approved drug chlorpromazine. *Int J Antimicrob Agents* **2021**, *57*, 106274, doi:10.1016/j.ijantimicag.2020.106274.
24. PFIZER, Pfizer initiates Phase 1 Study of novel oral antiviral therapeutic agent against SARS-CoV-2. **2021**.
25. *Halford, B.* Pfizer unveils its oral SARS-CoV-2 inhibitor **2021**.
26. Ou, X.; Liu, Y.; Lei, X.; Li, P.; Mi, D.; Ren, L.; Guo, L.; Guo, R.; Chen, T.; Hu, J.; et al. Characterization of spike glycoprotein of SARS-CoV-2 on virus entry and its immune cross-reactivity with SARS-CoV. *Nat Commun* **2020**, *11*, 1620, doi:10.1038/s41467-020-15562-9.
27. Mori, T.; O'Keefe, B.R.; Sowder, R.C.; Bringans, S.; Gardella, R.; Berg, S.; Cochran, P.; Turpin, J.A.; Buckheit, R.W.; McMahon, J.B.; et al. Isolation and characterization of griffithsin, a novel HIV-inactivating protein, from the red alga *Griffithsia* sp. *J Biol Chem* **2005**, *280*, 9345-9353, doi:10.1074/jbc.M411122200.
28. Shyr, Z.A.; Gorshkov, K.; Chen, C.Z.; Zheng, W. Drug Discovery Strategies for SARS-CoV-2. *J Pharmacol Exp Ther* **2020**, *375*, 127-138, doi:10.1124/jpet.120.000123.
29. Day, C.W.; Baric, R.; Cai, S.X.; Frieman, M.; Kumaki, Y.; Morrey, J.D.; Smee, D.F.; Barnard, D.L. A new mouse-adapted strain of SARS-CoV as a lethal model for evaluating antiviral agents in vitro and in vivo. *Virology* **2009**, *395*, 210-222, doi:10.1016/j.virol.2009.09.023.

30. Kumaki, Y.; Wandersee, M.K.; Smith, A.J.; Zhou, Y.; Simmons, G.; Nelson, N.M.; Bailey, K.W.; Vest, Z.G.; Li, J.K.; Chan, P.K.; et al. Inhibition of severe acute respiratory syndrome coronavirus replication in a lethal SARS-CoV BALB/c mouse model by stinging nettle lectin, *Urtica dioica* agglutinin. *Antiviral Res* **2011**, *90*, 22-32, doi:10.1016/j.antiviral.2011.02.003.
31. Yu, F.; Xiang, R.; Deng, X.; Wang, L.; Yu, Z.; Tian, S.; Liang, R.; Li, Y.; Ying, T.; Jiang, S. Receptor-binding domain-specific human neutralizing monoclonal antibodies against SARS-CoV and SARS-CoV-2. *Signal Transduct Target Ther* **2020**, *5*, 212, doi:10.1038/s41392-020-00318-0.
32. Chi, X.; Liu, X.; Wang, C.; Zhang, X.; Li, X.; Hou, J.; Ren, L.; Jin, Q.; Wang, J.; Yang, W. Humanized single domain antibodies neutralize SARS-CoV-2 by targeting the spike receptor binding domain. *Nat Commun* **2020**, *11*, 4528, doi:10.1038/s41467-020-18387-8.
33. Cao, Y.; Su, B.; Guo, X.; Sun, W.; Deng, Y.; Bao, L.; Zhu, Q.; Zhang, X.; Zheng, Y.; Geng, C.; et al. Potent Neutralizing Antibodies against SARS-CoV-2 Identified by High-Throughput Single-Cell Sequencing of Convalescent Patients' B Cells. *Cell* **2020**, *182*, 73-84.e16, doi:10.1016/j.cell.2020.05.025.
34. Li, W.; Drelich, A.; Martinez, D.R.; Gralinski, L.; Chen, C.; Sun, Z.; Liu, X.; Zhelev, D.; Zhang, L.; Peterson, E.C.; et al. Potent neutralization of SARS-CoV-2 in vitro and in an animal model by a human monoclonal antibody. *bioRxiv* **2020**, doi:10.1101/2020.05.13.093088.
35. Shi, R.; Shan, C.; Duan, X.; Chen, Z.; Liu, P.; Song, J.; Song, T.; Bi, X.; Han, C.; Wu, L.; et al. A human neutralizing antibody targets the receptor-binding site of SARS-CoV-2. *Nature* **2020**, *584*, 120-124, doi:10.1038/s41586-020-2381-y.
36. Wu, Y.; Wang, F.; Shen, C.; Peng, W.; Li, D.; Zhao, C.; Li, Z.; Li, S.; Bi, Y.; Yang, Y.; et al. A noncompeting pair of human neutralizing antibodies block COVID-19 virus binding to its receptor ACE2. *Science* **2020**, *368*, 1274-1278, doi:10.1126/science.abc2241.
37. Wrapp, D.; De Vlieger, D.; Corbett, K.S.; Torres, G.M.; Wang, N.; Van Breedam, W.; Roose, K.; van Schie, L.; Hoffmann, M.; Pöhlmann, S.; et al. Structural Basis for Potent Neutralization of Betacoronaviruses by Single-Domain Camelid Antibodies. *Cell* **2020**, *181*, 1004-1015.e1015, doi:10.1016/j.cell.2020.04.031.
38. Brouwer, P.J.M.; Caniels, T.G.; van der Straten, K.; Snitselaar, J.L.; Aldon, Y.; Bangaru, S.; Torres, J.L.; Okba, N.M.A.; Claireaux, M.; Kerster, G.; et al. Potent neutralizing antibodies from COVID-19 patients define multiple targets of vulnerability. *Science* **2020**, *369*, 643-650, doi:10.1126/science.abc5902.
39. Tai, W.; Zhang, X.; He, Y.; Jiang, S.; Du, L. Identification of SARS-CoV RBD-targeting monoclonal antibodies with cross-reactive or neutralizing activity against SARS-CoV-2. *Antiviral Res* **2020**, *179*, 104820, doi:10.1016/j.antiviral.2020.104820.
40. Wen, W.; Su, W.; Tang, H.; Le, W.; Zhang, X.; Zheng, Y.; Liu, X.; Xie, L.; Li, J.; Ye, J.; et al. Immune cell profiling of COVID-19 patients in the recovery stage by single-cell sequencing. *Cell Discov* **2020**, *6*, 31, doi:10.1038/s41421-020-0168-9.
41. Lachmann, H.J.; Kone-Paut, I.; Kuemmerle-Deschner, J.B.; Leslie, K.S.; Hachulla, E.; Quartier, P.; Gitton, X.; Widmer, A.; Patel, N.; Hawkins, P.N.; et al. Use of canakinumab in the cryopyrin-associated periodic syndrome. *N Engl J Med* **2009**, *360*, 2416-2425, doi:10.1056/NEJMoa0810787.

42. Gritti, G., F. Raimondi, D. Ripamonti, I. Riva, F. Landi, L. Alborghetti, M. Frigeni, M. Damiani, C. Mico, S. Faggioli, et al. Use of siltuximab in patients with COVID-19 pneumonia requiring ventilatory support. **2020**.
43. Burrage, D.R.; Koushesh, S.; Sofat, N. Immunomodulatory Drugs in the Management of SARS-CoV-2. *Front Immunol* **2020**, *11*, 1844, doi:10.3389/fimmu.2020.01844.
44. Mitchell, E.P. Corona Virus: Global Pandemic Causing World-Wide Shutdown. *J Natl Med Assoc* **2020**, *112*, 113-114, doi:10.1016/j.jnma.2020.03.015.
45. Chan, J.F.; Yuan, S.; Kok, K.H.; To, K.K.; Chu, H.; Yang, J.; Xing, F.; Liu, J.; Yip, C.C.; Poon, R.W.; et al. A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. *Lancet* **2020**, *395*, 514-523, doi:10.1016/S0140-6736(20)30154-9.
46. Lu, R.; Zhao, X.; Li, J.; Niu, P.; Yang, B.; Wu, H.; Wang, W.; Song, H.; Huang, B.; Zhu, N.; et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *Lancet* **2020**, *395*, 565-574, doi:10.1016/S0140-6736(20)30251-8.
47. Lai, M.M. Coronavirus: organization, replication and expression of genome. *Annu Rev Microbiol* **1990**, *44*, 303-333, doi:10.1146/annurev.mi.44.100190.001511.
48. Wu, F.; Zhao, S.; Yu, B.; Chen, Y.M.; Wang, W.; Song, Z.G.; Hu, Y.; Tao, Z.W.; Tian, J.H.; Pei, Y.Y.; et al. A new coronavirus associated with human respiratory disease in China. *Nature* **2020**, *579*, 265-269, doi:10.1038/s41586-020-2008-3.
49. Naqvi, A.A.T.; Fatima, K.; Mohammad, T.; Fatima, U.; Singh, I.K.; Singh, A.; Atif, S.M.; Hariprasad, G.; Hasan, G.M.; Hassan, M.I. Insights into SARS-CoV-2 genome, structure, evolution, pathogenesis and therapies: Structural genomics approach. *Biochim Biophys Acta Mol Basis Dis* **2020**, *1866*, 165878, doi:10.1016/j.bbadis.2020.165878.
50. Hoffmann, M.; Kleine-Weber, H.; Schroeder, S.; Krüger, N.; Herrler, T.; Erichsen, S.; Schiergens, T.S.; Herrler, G.; Wu, N.H.; Nitsche, A.; et al. SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. *Cell* **2020**, *181*, 271-280.e278, doi:10.1016/j.cell.2020.02.052.
51. Wrapp, D.; Wang, N.; Corbett, K.S.; Goldsmith, J.A.; Hsieh, C.L.; Abiona, O.; Graham, B.S.; McLellan, J.S. Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science* **2020**, *367*, 1260-1263, doi:10.1126/science.abb2507.
52. Ke, Z.; Oton, J.; Qu, K.; Cortese, M.; Zila, V.; McKeane, L.; Nakane, T.; Zivanov, J.; Neufeldt, C.J.; Cerikan, B.; et al. Structures and distributions of SARS-CoV-2 spike proteins on intact virions. *Nature* **2020**, *588*, 498-502, doi:10.1038/s41586-020-2665-2.
53. Qing, E.; Gallagher, T. SARS Coronavirus Redux. *Trends Immunol* **2020**, *41*, 271-273, doi:10.1016/j.it.2020.02.007.
54. Isa, F.; Gonzalez Ortiz, A.M.; Meyer, J.; Hamilton, J.D.; Olenchock, B.A.; Brackin, T.; Ganguly, S.; Forleo-Neto, E.; Faria, L.; Heirman, I.; et al. Effect of timing of casirivimab and imdevimab administration relative to mRNA-1273 COVID-19 vaccination on vaccine-induced SARS-CoV-2 neutralising antibody responses: a prospective, open-label, phase 2, randomised controlled trial. *Lancet Infect Dis* **2024**, doi:10.1016/S1473-3099(24)00421-3.
55. Walls, A.C.; Park, Y.J.; Tortorici, M.A.; Wall, A.; McGuire, A.T.; Velesler, D. Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. *Cell* **2020**, *181*, 281-292.e286, doi:10.1016/j.cell.2020.02.058.

56. Xia, S.; Yan, L.; Xu, W.; Agrawal, A.S.; Algaissi, A.; Tseng, C.K.; Wang, Q.; Du, L.; Tan, W.; Wilson, I.A.; et al. A pan-coronavirus fusion inhibitor targeting the HR1 domain of human coronavirus spike. *Sci Adv* **2019**, *5*, eaav4580, doi:10.1126/sciadv.aav4580.
57. Xu, J.; Zhao, S.; Teng, T.; Abdalla, A.E.; Zhu, W.; Xie, L.; Wang, Y.; Guo, X. Systematic Comparison of Two Animal-to-Human Transmitted Human Coronaviruses: SARS-CoV-2 and SARS-CoV. *Viruses* **2020**, *12*, doi:10.3390/v12020244.
58. Tang, T.; Bidon, M.; Jaimes, J.A.; Whittaker, G.R.; Daniel, S. Coronavirus membrane fusion mechanism offers a potential target for antiviral development. *Antiviral Res* **2020**, *178*, 104792, doi:10.1016/j.antiviral.2020.104792.
59. Neuman, B.W.; Buchmeier, M.J. Supramolecular Architecture of the Coronavirus Particle. *Adv Virus Res* **2016**, *96*, 1-27, doi:10.1016/bs.aivir.2016.08.005.
60. Cong, Y.; Ulasli, M.; Schepers, H.; Mauthe, M.; V'kovski, P.; Kriegenburg, F.; Thiel, V.; de Haan, C.A.M.; Reggiori, F. Nucleocapsid Protein Recruitment to Replication-Transcription Complexes Plays a Crucial Role in Coronaviral Life Cycle. *J Virol* **2020**, *94*, doi:10.1128/JVI.01925-19.
61. Fehr, A.R.; Perlman, S. Coronaviruses: an overview of their replication and pathogenesis. *Methods Mol Biol* **2015**, *1282*, 1-23, doi:10.1007/978-1-4939-2438-7\_1.
62. Fu, Z.; Huang, B.; Tang, J.; Liu, S.; Liu, M.; Ye, Y.; Liu, Z.; Xiong, Y.; Zhu, W.; Cao, D.; et al. The complex structure of GRL0617 and SARS-CoV-2 PLpro reveals a hot spot for antiviral drug discovery. *Nat Commun* **2021**, *12*, 488, doi:10.1038/s41467-020-20718-8.
63. Ullrich, S.; Nitsche, C. The SARS-CoV-2 main protease as drug target. *Bioorg Med Chem Lett* **2020**, *30*, 127377, doi:10.1016/j.bmcl.2020.127377.
64. Cho, E.; Rosa, M.; Anjum, R.; Mehmood, S.; Soban, M.; Mujtaba, M.; Bux, K.; Moin, S.T.; Tanweer, M.; Dantu, S.; et al. Dynamic Profiling of  $\beta$ -Coronavirus 3CL M<sup>PRO</sup> Protease Ligand-Binding Sites. *J Chem Inf Model* **2021**, *61*, 3058-3073, doi:10.1021/acs.jcim.1c00449.
65. Zhang, L.; Lin, D.; Sun, X.; Curth, U.; Drosten, C.; Sauerhering, L.; Becker, S.; Rox, K.; Hilgenfeld, R. Crystal structure of SARS-CoV-2 main protease provides a basis for design of improved  $\alpha$ -ketoamide inhibitors. *Science* **2020**, *368*, 409-412, doi:10.1126/science.abb3405.
66. Liang, J.; Karagiannis, C.; Pitsillou, E.; Darmawan, K.K.; Ng, K.; Hung, A.; Karagiannis, T.C. Site mapping and small molecule blind docking reveal a possible target site on the SARS-CoV-2 main protease dimer interface. *Comput Biol Chem* **2020**, *89*, 107372, doi:10.1016/j.compbiolchem.2020.107372.
67. Suárez, D.; Díaz, N. SARS-CoV-2 Main Protease: A Molecular Dynamics Study. *J Chem Inf Model* **2020**, *60*, 5815-5831, doi:10.1021/acs.jcim.0c00575.
68. El-Baba, T.J.; Lutowski, C.A.; Kantsadi, A.L.; Malla, T.R.; John, T.; Mikhailov, V.; Bolla, J.R.; Schofield, C.J.; Zitzmann, N.; Vakonakis, I.; et al. Allosteric Inhibition of the SARS-CoV-2 Main Protease: Insights from Mass Spectrometry Based Assays\*. *Angew Chem Int Ed Engl* **2020**, *59*, 23544-23548, doi:10.1002/anie.202010316.
69. El Ahdab, D.; Lagardère, L.; Inizan, T.J.; Célerse, F.; Liu, C.; Adjoua, O.; Jolly, L.H.; Gresh, N.; Hobaika, Z.; Ren, P.; et al. Interfacial Water Many-Body Effects Drive Structural Dynamics and Allosteric Interactions in SARS-CoV-2 Main Protease Dimerization Interface. *J Phys Chem Lett* **2021**, *12*, 6218-6226, doi:10.1021/acs.jpcllett.1c01460.

70. Jin, Z.; Du, X.; Xu, Y.; Deng, Y.; Liu, M.; Zhao, Y.; Zhang, B.; Li, X.; Zhang, L.; Peng, C.; et al. Structure of M<sup>pro</sup> from SARS-CoV-2 and discovery of its inhibitors. *Nature* **2020**, *582*, 289-293, doi:10.1038/s41586-020-2223-y.
71. Arafet, K.; Serrano-Aparicio, N.; Lodola, A.; Mulholland, A.J.; González, F.V.; Świderek, K.; Moliner, V. Mechanism of inhibition of SARS-CoV-2 M<sup>pro</sup> by N3 peptidyl Michael acceptor explained by QM/MM simulations and design of new derivatives with tunable chemical reactivity. *Chem Sci* **2020**, *12*, 1433-1444, doi:10.1039/d0sc06195f.
72. Ramos-Guzmán, C.A.; Ruiz-Pernía, J.J.; Tuñón, I. Unraveling the SARS-CoV-2 Main Protease Mechanism Using Multiscale Methods. *ACS Catal* **2020**, *10*, 12544-12554, doi:10.1021/acscatal.0c03420.
73. Świderek, K.; Moliner, V. Revealing the molecular mechanisms of proteolysis of SARS-CoV-2 M. *Chem Sci* **2020**, *11*, 10626-10630, doi:10.1039/d0sc02823a.
74. Muramatsu, T.; Takemoto, C.; Kim, Y.T.; Wang, H.; Nishii, W.; Terada, T.; Shirouzu, M.; Yokoyama, S. SARS-CoV 3CL protease cleaves its C-terminal autoprocessing site by novel subsite cooperativity. *Proc Natl Acad Sci U S A* **2016**, *113*, 12997-13002, doi:10.1073/pnas.1601327113.
75. Fan, K.; Ma, L.; Han, X.; Liang, H.; Wei, P.; Liu, Y.; Lai, L. The substrate specificity of SARS coronavirus 3C-like proteinase. *Biochem Biophys Res Commun* **2005**, *329*, 934-940, doi:10.1016/j.bbrc.2005.02.061.
76. La Monica, G.; Bono, A.; Lauria, A.; Martorana, A. Targeting SARS-CoV-2 Main Protease for Treatment of COVID-19: Covalent Inhibitors Structure-Activity Relationship Insights and Evolution Perspectives. *J Med Chem* **2022**, doi:10.1021/acs.jmedchem.2c01005.
77. Goyal, B.; Goyal, D. Targeting the Dimerization of the Main Protease of Coronaviruses: A Potential Broad-Spectrum Therapeutic Strategy. *ACS Comb Sci* **2020**, *22*, 297-305, doi:10.1021/acscombsci.0c00058.
78. Sztain, T.; Amaro, R.; McCammon, J.A. Elucidation of Cryptic and Allosteric Pockets within the SARS-CoV-2 Main Protease. *J Chem Inf Model* **2021**, *61*, 3495-3501, doi:10.1021/acs.jcim.1c00140.
79. Carli, M.; Sormani, G.; Rodriguez, A.; Laio, A. Candidate Binding Sites for Allosteric Inhibition of the SARS-CoV-2 Main Protease from the Analysis of Large-Scale Molecular Dynamics Simulations. *J Phys Chem Lett* **2021**, *12*, 65-72, doi:10.1021/acs.jpcllett.0c03182.
80. Dubanevics, I.; McLeish, T.C.B. Computational analysis of dynamic allostery and control in the SARS-CoV-2 main protease. *J R Soc Interface* **2021**, *18*, 20200591, doi:10.1098/rsif.2020.0591.
81. Alzyoud, L.; Ghattas, M.A.; Atatreh, N. Allosteric Binding Sites of the SARS-CoV-2 Main Protease: Potential Targets for Broad-Spectrum Anti-Coronavirus Agents. *Drug Des Devel Ther* **2022**, *16*, 2463-2478, doi:10.2147/DDDT.S370574.
82. Aljoundi, A.; Bji, I.; El Rashedy, A.; Soliman, M.E.S. Covalent Versus Non-covalent Enzyme Inhibition: Which Route Should We Take? A Justification of the Good and Bad from Molecular Modelling Perspective. *Protein J* **2020**, *39*, 97-105, doi:10.1007/s10930-020-09884-2.
83. Günther, S.; Reinke, P.Y.A.; Fernández-García, Y.; Lieske, J.; Lane, T.J.; Ginn, H.M.; Koua, F.H.M.; Ehart, C.; Ewert, W.; Oberthuer, D.; et al. X-ray screening identifies active site and allosteric inhibitors of SARS-CoV-2 main protease. *Science* **2021**, *372*, 642-646, doi:10.1126/science.abf7945.

84. Dolman, M.E.; Poon, E.; Ebus, M.E.; den Hartog, I.J.; van Noesel, C.J.; Jamin, Y.; Hallsworth, A.; Robinson, S.P.; Petrie, K.; Sparidans, R.W.; et al. Cyclin-Dependent Kinase Inhibitor AT7519 as a Potential Drug for MYCN-Dependent Neuroblastoma. *Clin Cancer Res* **2015**, *21*, 5100-5109, doi:10.1158/1078-0432.CCR-15-0313.
85. Chaves, O.A.; Sacramento, C.Q.; Fintelman-Rodrigues, N.; Temerozo, J.R.; Pereira-Dutra, F.; Mizurini, D.M.; Monteiro, R.Q.; Vazquez, L.; Bozza, P.T.; Castro-Faria-Neto, H.C.; et al. Apixaban, an orally available anticoagulant, inhibits SARS-CoV-2 replication and its major protease in a non-competitive way. *J Mol Cell Biol* **2022**, *14*, doi:10.1093/jmcb/mjac039.
86. Chaves, O.A.; Lima, C.R.; Fintelman-Rodrigues, N.; Sacramento, C.Q.; de Freitas, C.S.; Vazquez, L.; Temerozo, J.R.; Rocha, M.E.N.; Dias, S.S.G.; Carels, N.; et al. Agathisflavone, a natural biflavonoid that inhibits SARS-CoV-2 replication by targeting its proteases. *Int J Biol Macromol* **2022**, *222*, 1015-1026, doi:10.1016/j.ijbiomac.2022.09.204.
87. Unoh, Y.; Uehara, S.; Nakahara, K.; Nobori, H.; Yamatsu, Y.; Yamamoto, S.; Maruyama, Y.; Taoda, Y.; Kasamatsu, K.; Suto, T.; et al. Discovery of S-217622, a Noncovalent Oral SARS-CoV-2 3CL Protease Inhibitor Clinical Candidate for Treating COVID-19. *J Med Chem* **2022**, *65*, 6499-6512, doi:10.1021/acs.jmedchem.2c00117.
88. Lin, M.; Zeng, X.; Duan, Y.; Yang, Z.; Ma, Y.; Yang, H.; Yang, X.; Liu, X. Molecular mechanism of ensitrelvir inhibiting SARS-CoV-2 main protease and its variants. *Commun Biol* **2023**, *6*, 694, doi:10.1038/s42003-023-05071-y.
89. Mukae, H.; Yotsuyanagi, H.; Ohmagari, N.; Doi, Y.; Imamura, T.; Sonoyama, T.; Fukuhara, T.; Ichihashi, G.; Sanaki, T.; Baba, K.; et al. A Randomized Phase 2/3 Study of Ensitrelvir, a Novel Oral SARS-CoV-2 3C-Like Protease Inhibitor, in Japanese Patients with Mild-to-Moderate COVID-19 or Asymptomatic SARS-CoV-2 Infection: Results of the Phase 2a Part. *Antimicrob Agents Chemother* **2022**, *66*, e0069722, doi:10.1128/aac.00697-22.
90. Anton, D.B.; Galvez Bulhões Pedreira, J.; Zvirtes, M.L.; Laufer, S.A.; Ducati, R.G.; Goettert, M.; Saraiva Macedo Timmers, L.F. Targeting SARS-CoV-2 Main Protease (MPro) with Kinase Inhibitors: A Promising Approach for Discovering Antiviral and Anti-inflammatory Molecules against SARS-CoV-2. *J Chem Inf Model* **2023**, *63*, 4138-4146, doi:10.1021/acs.jcim.3c00324.
91. Selvaraj, V.; Finn, A.; Lal, A.; Khan, M.S.; Dapaah-Afriyie, K.; Carino, G.P. Baricitinib in hospitalised patients with COVID-19: A meta-analysis of randomised controlled trials. *EClinicalMedicine* **2022**, *49*, 101489, doi:10.1016/j.eclinm.2022.101489.
92. Drayman, N.; DeMarco, J.K.; Jones, K.A.; Azizi, S.A.; Froggatt, H.M.; Tan, K.; Maltseva, N.I.; Chen, S.; Nicolaescu, V.; Dvorkin, S.; et al. Masitinib is a broad coronavirus 3CL inhibitor that blocks replication of SARS-CoV-2. *Science* **2021**, *373*, 931-936, doi:10.1126/science.abg5827.
93. Ketabforoush, A.H.M.E.; Chegini, R.; Barati, S.; Tahmasebi, F.; Moghiseh, B.; Joghataei, M.T.; Faghihi, F.; Azedi, F. Masitinib: The promising actor in the next season of the Amyotrophic Lateral Sclerosis treatment series. *Biomed Pharmacother* **2023**, *160*, 114378, doi:10.1016/j.biopha.2023.114378.
94. Han, S.H.; Goins, C.M.; Arya, T.; Shin, W.J.; Maw, J.; Hooper, A.; Sonawane, D.P.; Porter, M.R.; Bannister, B.E.; Crouch, R.D.; et al. Structure-Based Optimization of ML300-Derived, Noncovalent Inhibitors Targeting the Severe Acute Respiratory



- Syndrome Coronavirus 3CL Protease (SARS-CoV-2 3CL<sup>PRO</sup>). *J Med Chem* **2022**, *65*, 2880-2904, doi:10.1021/acs.jmedchem.1c00598.
95. Chen, R.; Gao, Y.; Liu, H.; Li, H.; Chen, W.; Ma, J. Advances in research on 3C-like protease (3CLpro) inhibitors against SARS-CoV-2 since 2020. *RSC Med Chem* **2023**, *14*, 9-21, doi:10.1039/d2md00344a.
  96. Du, A.; Zheng, R.; Disoma, C.; Li, S.; Chen, Z.; Liu, P.; Zhou, Y.; Shen, Y.; Liu, S.; Zhang, Y.; et al. Epigallocatechin-3-gallate, an active ingredient of Traditional Chinese Medicines, inhibits the 3CLpro activity of SARS-CoV-2. *Int J Biol Macromol* **2021**, *176*, 1-12, doi:10.1016/j.ijbiomac.2021.02.012.
  97. Abian, O.; Ortega-Alarcon, D.; Jimenez-Alesanco, A.; Ceballos-Laita, L.; Vega, S.; Reyburn, H.T.; Rizzuti, B.; Velazquez-Campoy, A. Structural stability of SARS-CoV-2 3CLpro and identification of quercetin as an inhibitor by experimental screening. *Int J Biol Macromol* **2020**, *164*, 1693-1703, doi:10.1016/j.ijbiomac.2020.07.235.
  98. Rizzuti, B.; Grande, F.; Conforti, F.; Jimenez-Alesanco, A.; Ceballos-Laita, L.; Ortega-Alarcon, D.; Vega, S.; Reyburn, H.T.; Abian, O.; Velazquez-Campoy, A. Rutin Is a Low Micromolar Inhibitor of SARS-CoV-2 Main Protease 3CLpro: Implications for Drug Design of Quercetin Analogs. *Biomedicines* **2021**, *9*, doi:10.3390/biomedicines9040375.
  99. Bahun, M.; Jukić, M.; Oblak, D.; Kranjc, L.; Bajc, G.; Butala, M.; Bozovičar, K.; Bratkovič, T.; Podlipnik, Č.; Poklar Ulrih, N. Inhibition of the SARS-CoV-2 3CL<sup>PRO</sup> main protease by plant polyphenols. *Food Chem* **2022**, *373*, 131594, doi:10.1016/j.foodchem.2021.131594.
  100. Liao, Q.; Chen, Z.; Tao, Y.; Zhang, B.; Wu, X.; Yang, L.; Wang, Q.; Wang, Z. An integrated method for optimized identification of effective natural inhibitors against SARS-CoV-2 3CLpro. *Sci Rep* **2021**, *11*, 22796, doi:10.1038/s41598-021-02266-3.
  101. Di Pierro, F.; Iqtadar, S.; Khan, A.; Ullah Mumtaz, S.; Masud Chaudhry, M.; Bertuccioli, A.; Derosa, G.; Maffioli, P.; Togni, S.; Riva, A.; et al. Potential Clinical Benefits of Quercetin in the Early Stage of COVID-19: Results of a Second, Pilot, Randomized, Controlled and Open-Label Clinical Trial. *Int J Gen Med* **2021**, *14*, 2807-2816, doi:10.2147/IJGM.S318949.
  102. Di Pierro, F.; Khan, A.; Iqtadar, S.; Mumtaz, S.U.; Chaudhry, M.N.A.; Bertuccioli, A.; Derosa, G.; Maffioli, P.; Togni, S.; Riva, A.; et al. Quercetin as a possible complementary agent for early-stage COVID-19: Concluding results of a randomized clinical trial. *Front Pharmacol* **2022**, *13*, 1096853, doi:10.3389/fphar.2022.1096853.
  103. Imran, M.; Thabet, H.K.; Alaql, S.I.; Alzahrani, A.R.; Abida, A.; Alshammari, M.K.; Kamal, M.; Diwan, A.; Asdaq, S.M.B.; Alshehri, S. The Therapeutic and Prophylactic Potential of Quercetin against COVID-19: An Outlook on the Clinical Studies, Inventive Compositions, and Patent Literature. *Antioxidants (Basel)* **2022**, *11*, doi:10.3390/antiox11050876.
  104. Yang, M.; Wei, J.; Huang, T.; Lei, L.; Shen, C.; Lai, J.; Liu, L.; Yang, Y.; Liu, G.; Liu, Y. Resveratrol inhibits the replication of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) in cultured Vero cells. *Phytother Res* **2021**, *35*, 1127-1129, doi:10.1002/ptr.6916.
  105. Pasquereau, S.; Nehme, Z.; Haidar Ahmad, S.; Daouad, F.; Van Assche, J.; Wallet, C.; Schwartz, C.; Rohr, O.; Morot-Bizot, S.; Herbein, G. Resveratrol Inhibits HCoV-229E and SARS-CoV-2 Coronavirus Replication In Vitro. *Viruses* **2021**, *13*, doi:10.3390/v13020354.

106. McCreary, M.R.; Schnell, P.M.; Rhoda, D.A. Randomized double-blind placebo-controlled proof-of-concept trial of resveratrol for outpatient treatment of mild coronavirus disease (COVID-19). *Sci Rep* **2022**, *12*, 10978, doi:10.1038/s41598-022-13920-9.
107. Marín-Palma, D.; Tabares-Guevara, J.H.; Zapata-Cardona, M.I.; Flórez-Álvarez, L.; Yepes, L.M.; Rugeles, M.T.; Zapata-Builes, W.; Hernandez, J.C.; Taborda, N.A. Curcumin Inhibits In Vitro SARS-CoV-2 Infection In Vero E6 Cells through Multiple Antiviral Mechanisms. *Molecules* **2021**, *26*, doi:10.3390/molecules26226900.
108. Lauria, A.; Mannino, S.; Gentile, C.; Mannino, G.; Martorana, A.; Peri, D. DRUDIT: web-based DRUGs Discovery Tools to design small molecules as modulators of biological targets. *Bioinformatics* **2020**, *36*, 1562-1569, doi:10.1093/bioinformatics/btz783.
109. Martorana, A.; Gentile, C.; Perricone, U.; Piccionello, A.P.; Bartolotta, R.; Terenzi, A.; Pace, A.; Mingoia, F.; Almerico, A.M.; Lauria, A. Synthesis, antiproliferative activity, and in silico insights of new 3-benzoylamino-benzo[b]thiophene derivatives. *Eur J Med Chem* **2015**, *90*, 537-546, doi:10.1016/j.ejmech.2014.12.002.
110. Lauria, A.; Gentile, C.; Mingoia, F.; Palumbo Piccionello, A.; Bartolotta, R.; Delisi, R.; Buscemi, S.; Martorana, A. Design, synthesis, and biological evaluation of a new class of benzo[b]furan derivatives as antiproliferative agents, with in silico predicted antitubulin activity. *Chem Biol Drug Des* **2018**, *91*, 39-49, doi:10.1111/cbdd.13052.
111. Daina, A.; Michielin, O.; Zoete, V. SwissADME: a free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules. *Sci Rep* **2017**, *7*, 42717, doi:10.1038/srep42717.
112. Baell, J.B.; Holloway, G.A. New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays. *J Med Chem* **2010**, *53*, 2719-2740, doi:10.1021/jm901137j.
113. Lipinski, C.A.; Lombardo, F.; Dominy, B.W.; Feeney, P.J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv Drug Deliv Rev* **2001**, *46*, 3-26, doi:10.1016/s0169-409x(00)00129-0.
114. Veber, D.F.; Johnson, S.R.; Cheng, H.Y.; Smith, B.R.; Ward, K.W.; Kopple, K.D. Molecular properties that influence the oral bioavailability of drug candidates. *J Med Chem* **2002**, *45*, 2615-2623, doi:10.1021/jm020017n.
115. Egan, W.J.; Merz, K.M.; Baldwin, J.J. Prediction of drug absorption using multivariate statistics. *J Med Chem* **2000**, *43*, 3867-3877, doi:10.1021/jm000292e.
116. Zhao, Y.; Fang, C.; Zhang, Q.; Zhang, R.; Zhao, X.; Duan, Y.; Wang, H.; Zhu, Y.; Feng, L.; Zhao, J.; et al. Crystal structure of SARS-CoV-2 main protease in complex with protease inhibitor PF-07321332. *Protein Cell* **2021**, doi:10.1007/s13238-021-00883-2.
117. Sharma, A.; Kaliya, K.; Maurya, S.K. Recent Advances in the Discovery of Potent Proteases Inhibitors Targeting the SARS Coronaviruses. *Curr Top Med Chem* **2021**, *21*, 307-328, doi:10.2174/1568026620999201111160035.
118. Shindo, N.; Ojida, A. Recent progress in covalent warheads for in vivo targeting of endogenous proteins. *Bioorg Med Chem* **2021**, *47*, 116386, doi:10.1016/j.bmc.2021.116386.
119. Liu, Y.; Lv, S.; Peng, L.; Xie, C.; Gao, L.; Sun, H.; Lin, L.; Ding, K.; Li, Z. Development and application of novel electrophilic warheads in target identification and drug discovery. *Biochem Pharmacol* **2021**, *190*, 114636, doi:10.1016/j.bcp.2021.114636.

120. Gehringer, M.; Laufer, S.A. Emerging and Re-Emerging Warheads for Targeted Covalent Inhibitors: Applications in Medicinal Chemistry and Chemical Biology. *J Med Chem* **2019**, *62*, 5673-5724, doi:10.1021/acs.jmedchem.8b01153.
121. Yang, H.; Xie, W.; Xue, X.; Yang, K.; Ma, J.; Liang, W.; Zhao, Q.; Zhou, Z.; Pei, D.; Ziebuhr, J.; et al. Design of wide-spectrum inhibitors targeting coronavirus main proteases. *PLoS Biol* **2005**, *3*, e324, doi:10.1371/journal.pbio.0030324.
122. Dai, W.; Zhang, B.; Jiang, X.M.; Su, H.; Li, J.; Zhao, Y.; Xie, X.; Jin, Z.; Peng, J.; Liu, F.; et al. Structure-based design of antiviral drug candidates targeting the SARS-CoV-2 main protease. *Science* **2020**, *368*, 1331-1335, doi:10.1126/science.abb4489.
123. Ramos-Guzmán, C.A.; Ruiz-Pernía, J.J.; Tuñón, I. Multiscale Simulations of SARS-CoV-2 3CL Protease Inhibition with Aldehyde Derivatives. Role of Protein and Inhibitor Conformational Changes in the Reaction Mechanism. *ACS Catal* **2021**, *11*, 4157-4168, doi:10.1021/acscatal.0c05522.
124. Dai, W.; Jochmans, D.; Xie, H.; Yang, H.; Li, J.; Su, H.; Chang, D.; Wang, J.; Peng, J.; Zhu, L.; et al. Design, Synthesis, and Biological Evaluation of Peptidomimetic Aldehydes as Broad-Spectrum Inhibitors against Enterovirus and SARS-CoV-2. *J Med Chem* **2021**, doi:10.1021/acs.jmedchem.0c02258.
125. Kim, Y.; Lovell, S.; Tiew, K.C.; Mandadapu, S.R.; Alliston, K.R.; Battaile, K.P.; Groutas, W.C.; Chang, K.O. Broad-spectrum antivirals against 3C or 3C-like proteases of picornaviruses, noroviruses, and coronaviruses. *J Virol* **2012**, *86*, 11754-11762, doi:10.1128/JVI.01348-12.
126. Vuong, W.; Khan, M.B.; Fischer, C.; Arutyunova, E.; Lamer, T.; Shields, J.; Saffran, H.A.; McKay, R.T.; van Belkum, M.J.; Joyce, M.A.; et al. Feline coronavirus drug inhibits the main protease of SARS-CoV-2 and blocks virus replication. *Nat Commun* **2020**, *11*, 4282, doi:10.1038/s41467-020-18096-2.
127. Ma, C.; Sacco, M.D.; Hurst, B.; Townsend, J.A.; Hu, Y.; Szeto, T.; Zhang, X.; Tarbet, B.; Marty, M.T.; Chen, Y.; et al. Boceprevir, GC-376, and calpain inhibitors II, XII inhibit SARS-CoV-2 viral replication by targeting the viral main protease. *Cell Res* **2020**, *30*, 678-692, doi:10.1038/s41422-020-0356-z.
128. Hu, Y.; Ma, C.; Szeto, T.; Hurst, B.; Tarbet, B.; Wang, J. Boceprevir, Calpain Inhibitors II and XII, and GC-376 Have Broad-Spectrum Antiviral Activity against Coronaviruses. *ACS Infect Dis* **2021**, *7*, 586-597, doi:10.1021/acsinfecdis.0c00761.
129. Simmons, G.; Gosalia, D.N.; Rennekamp, A.J.; Reeves, J.D.; Diamond, S.L.; Bates, P. Inhibitors of cathepsin L prevent severe acute respiratory syndrome coronavirus entry. *Proc Natl Acad Sci U S A* **2005**, *102*, 11876-11881, doi:10.1073/pnas.0505577102.
130. Kumar, P.; Ratia, K.M.; Richner, J.M.; Thatcher, G.R.J.; Kadam, R.; Smieszek, S.P.; Przychodzen, B.P.; Koprivica, V.; Birznieks, G.; Polymeropoulos, M.H.; et al. Dual Inhibition of Cathepsin L and 3CL-Pro by GC-376 Constrains SARS Cov2 Infection Including Omicron Variant. *bioRxiv* **2022**, 2022.2002.2009.479835, doi:10.1101/2022.02.09.479835.
131. Cáceres, C.J.; Cardenas-Garcia, S.; Carnaccini, S.; Seibert, B.; Rajao, D.S.; Wang, J.; Perez, D.R. Efficacy of GC-376 against SARS-CoV-2 virus infection in the K18 hACE2 transgenic mouse model. *Sci Rep* **2021**, *11*, 9609, doi:10.1038/s41598-021-89013-w.
132. Vuong, W.; Fischer, C.; Khan, M.B.; van Belkum, M.J.; Lamer, T.; Willoughby, K.D.; Lu, J.; Arutyunova, E.; Joyce, M.A.; Saffran, H.A.; et al. Improved SARS-CoV-2 M<sup>pro</sup> inhibitors based on feline antiviral drug GC376: Structural enhancements, increased

- solubility, and micellar studies. *Eur J Med Chem* **2021**, *222*, 113584, doi:10.1016/j.ejmech.2021.113584.
133. Sacco, M.D.; Ma, C.; Lagarias, P.; Gao, A.; Townsend, J.A.; Meng, X.; Dube, P.; Zhang, X.; Hu, Y.; Kitamura, N.; et al. Structure and inhibition of the SARS-CoV-2 main protease reveal strategy for developing dual inhibitors against M<sup>pro</sup> and cathepsin L. *Sci Adv* **2020**, *6*, doi:10.1126/sciadv.abe0751.
  134. Liu, H.; Iketani, S.; Zask, A.; Khanizeman, N.; Bednarova, E.; Forouhar, F.; Fowler, B.; Hong, S.J.; Mohri, H.; Nair, M.S.; et al. Development of optimized drug-like small molecule inhibitors of the SARS-CoV-2 3CL protease for treatment of COVID-19. *Nat Commun* **2022**, *13*, 1891, doi:10.1038/s41467-022-29413-2.
  135. Qiao, J.; Li, Y.S.; Zeng, R.; Liu, F.L.; Luo, R.H.; Huang, C.; Wang, Y.F.; Zhang, J.; Quan, B.; Shen, C.; et al. SARS-CoV-2 M<sup>pro</sup> inhibitors with antiviral activity in a transgenic mouse model. *Science* **2021**, *371*, 1374-1378, doi:10.1126/science.abf1611.
  136. Luo, D.; Tong, J.B.; Zhang, X.; Xiao, X.C.; Bian, S. Computational strategies towards developing novel SARS-CoV-2 M<sup>pro</sup> inhibitors against COVID-19. *J Mol Struct* **2021**, 131378, doi:10.1016/j.molstruc.2021.131378.
  137. Xia, Z.; Sacco, M.; Hu, Y.; Ma, C.; Meng, X.; Zhang, F.; Szeto, T.; Xiang, Y.; Chen, Y.; Wang, J. Rational Design of Hybrid SARS-CoV-2 Main Protease Inhibitors Guided by the Superimposed Cocrystal Structures with the Peptidomimetic Inhibitors GC-376, Telaprevir, and Boceprevir. *ACS Pharmacol Transl Sci* **2021**, *4*, 1408-1421, doi:10.1021/acspsci.1c00099.
  138. Yang, K.S.; Ma, X.R.; Ma, Y.; Alugubelli, Y.R.; Scott, D.A.; Vatansever, E.C.; Drelich, A.K.; Sankaran, B.; Geng, Z.Z.; Blankenship, L.R.; et al. A Quick Route to Multiple Highly Potent SARS-CoV-2 Main Protease Inhibitors. *ChemMedChem* **2021**, *16*, 942-948, doi:10.1002/cmdc.202000924.
  139. Ma, X.R.; Alugubelli, Y.R.; Ma, Y.; Vatansever, E.C.; Scott, D.A.; Qiao, Y.; Yu, G.; Xu, S.; Liu, W.R. MPI8 is Potent against SARS-CoV-2 by Inhibiting Dually and Selectively the SARS-CoV-2 Main Protease and the Host Cathepsin L. *ChemMedChem* **2022**, *17*, e202100456, doi:10.1002/cmdc.202100456.
  140. Costanzi, E.; Kuzikov, M.; Esposito, F.; Albani, S.; Demitri, N.; Giabbai, B.; Camasta, M.; Tramontano, E.; Rossetti, G.; Zaliani, A.; et al. Structural and Biochemical Analysis of the Dual Inhibition of MG-132 against SARS-CoV-2 Main Protease (M<sup>pro</sup>/3CL<sup>pro</sup>) and Human Cathepsin-L. *Int J Mol Sci* **2021**, *22*, doi:10.3390/ijms222111779.
  141. Kuzikov, M.; Costanzi, E.; Reinshagen, J.; Esposito, F.; Vangeel, L.; Wolf, M.; Ellinger, B.; Claussen, C.; Geisslinger, G.; Corona, A.; et al. Identification of Inhibitors of SARS-CoV-2 3CL-Pro Enzymatic Activity Using a Small Molecule in Vitro Repurposing Screen. *ACS Pharmacology and Translational Science* **2021**, *4*, 1096-1110, doi:10.1021/acspsci.0c00216.
  142. Hoffman, R.L.; Kania, R.S.; Brothers, M.A.; Davies, J.F.; Ferre, R.A.; Gajiwala, K.S.; He, M.; Hogan, R.J.; Kozminski, K.; Li, L.Y.; et al. Discovery of Ketone-Based Covalent Inhibitors of Coronavirus 3CL Proteases for the Potential Therapeutic Treatment of COVID-19. *J Med Chem* **2020**, *63*, 12725-12747, doi:10.1021/acs.jmedchem.0c01063.
  143. de Vries, M.; Mohamed, A.S.; Prescott, R.A.; Valero-Jimenez, A.M.; Desvignes, L.; O'Connor, R.; Steppan, C.; Devlin, J.C.; Ivanova, E.; Herrera, A.; et al. A comparative analysis of SARS-CoV-2 antivirals characterizes 3CL<sup>pro</sup> inhibitor PF-00835231 as a Potential New Treatment for COVID-19. *J Virol* **2021**, doi:10.1128/JVI.01819-20.

144. Boras, B.; Jones, R.M.; Anson, B.J.; Arenson, D.; Aschenbrenner, L.; Bakowski, M.A.; Beutler, N.; Binder, J.; Chen, E.; Eng, H.; et al. Preclinical characterization of an intravenous coronavirus 3CL protease inhibitor for the potential treatment of COVID19. *Nat Commun* **2021**, *12*, 6055, doi:10.1038/s41467-021-26239-2.
145. Bai, B.; Belovodskiy, A.; Hena, M.; Kandadai, A.S.; Joyce, M.A.; Saffran, H.A.; Shields, J.A.; Khan, M.B.; Arutyunova, E.; Lu, J.; et al. Peptidomimetic  $\alpha$ -Acyloxymethylketone Warheads with Six-Membered Lactam P1 Glutamine Mimic: SARS-CoV-2 3CL Protease Inhibition, Coronavirus Antiviral Activity, and *in Vitro* Biological Stability. *J Med Chem* **2021**, doi:10.1021/acs.jmedchem.1c00616.
146. Konno, S.; Kobayashi, K.; Senda, M.; Funai, Y.; Seki, Y.; Tamai, I.; Schäkel, L.; Sakata, K.; Pillaiyar, T.; Taguchi, A.; et al. 3CL Protease Inhibitors with an Electrophilic Arylketone Moiety as Anti-SARS-CoV-2 Agents. *J Med Chem* **2021**, doi:10.1021/acs.jmedchem.1c00665.
147. Hattori, S.I.; Higashi-Kuwata, N.; Hayashi, H.; Allu, S.R.; Raghavaiah, J.; Bulut, H.; Das, D.; Anson, B.J.; Lendy, E.K.; Takamatsu, Y.; et al. A small molecule compound with an indole moiety inhibits the main protease of SARS-CoV-2 and blocks virus replication. *Nat Commun* **2021**, *12*, 668, doi:10.1038/s41467-021-20900-6.
148. Zhang, L.; Lin, D.; Kusov, Y.; Nian, Y.; Ma, Q.; Wang, J.; von Brunn, A.; Leyssen, P.; Lanko, K.; Neyts, J.; et al.  $\alpha$ -Ketoamides as Broad-Spectrum Inhibitors of Coronavirus and Enterovirus Replication: Structure-Based Design, Synthesis, and Activity Assessment. *J Med Chem* **2020**, *63*, 4562-4578, doi:10.1021/acs.jmedchem.9b01828.
149. Banerjee, S. An insight into the interaction between  $\alpha$ -ketoamide-based inhibitor and coronavirus main protease: A detailed *in silico* study. *Biophys Chem* **2021**, *269*, 106510, doi:10.1016/j.bpc.2020.106510.
150. Mondal, D.; Warshel, A. Exploring the Mechanism of Covalent Inhibition: Simulating the Binding Free Energy of  $\alpha$ -Ketoamide Inhibitors of the Main Protease of SARS-CoV-2. *Biochemistry* **2020**, *59*, 4601-4608, doi:10.1021/acs.biochem.0c00782.
151. Liang, J.; Pitsillou, E.; Karagiannis, C.; Darmawan, K.K.; Ng, K.; Hung, A.; Karagiannis, T.C. Interaction of the prototypical  $\alpha$ -ketoamide inhibitor with the SARS-CoV-2 main protease active site *in silico*: Molecular dynamic simulations highlight the stability of the ligand-protein complex. *Comput Biol Chem* **2020**, *87*, 107292, doi:10.1016/j.compbiolchem.2020.107292.
152. Oubahmane, M.; Hdoufane, I.; Bjiij, I.; Jerves, C.; Villemin, D.; Cherqaoui, D. COVID-19: *In silico* identification of potent  $\alpha$ -ketoamide inhibitors targeting the main protease of the SARS-CoV-2. *J Mol Struct* **2021**, *1244*, 130897, doi:10.1016/j.molstruc.2021.130897.
153. Frecer, V.; Miertus, S. Antiviral agents against COVID-19: Structure-based design of specific peptidomimetic inhibitors of SARS-CoV-2 main protease. *RSC Advances* **2020**, *10*, 40244-40263, doi:10.1039/d0ra08304f.
154. Oerlemans, R.; Ruiz-Moreno, A.J.; Cong, Y.; Dinesh Kumar, N.; Velasco-Velazquez, M.A.; Neochoritis, C.G.; Smith, J.; Reggiori, F.; Groves, M.R.; Dömling, A. Repurposing the HCV NS3-4A protease drug boceprevir as COVID-19 therapeutics. *RSC Med Chem* **2020**, *12*, 370-379, doi:10.1039/d0md00367k.
155. Elzupir, A.O. Inhibition of SARS-CoV-2 main protease 3CL<sup>pro</sup> by means of  $\alpha$ -ketoamide and pyridone-containing pharmaceuticals using *in silico* molecular docking. *J Mol Struct* **2020**, *1222*, 128878, doi:10.1016/j.molstruc.2020.128878.

156. Kneller, D.W.; Phillips, G.; Weiss, K.L.; Zhang, Q.; Coates, L.; Kovalevsky, A. Direct Observation of Protonation State Modulation in SARS-CoV-2 Main Protease upon Inhibitor Binding with Neutron Crystallography. *J Med Chem* **2021**, *64*, 4991-5000, doi:10.1021/acs.jmedchem.1c00058.
157. Ramos-Guzmán, C.A.; Ruiz-Pernía, J.J.; Tuñón, I. A microscopic description of SARS-CoV-2 main protease inhibition with Michael acceptors. Strategies for improving inhibitor design. *Chem Sci* **2021**, *12*, 3489-3496, doi:10.1039/d0sc04978f.
158. Iketani, S.; Forouhar, F.; Liu, H.; Hong, S.J.; Lin, F.Y.; Nair, M.S.; Zask, A.; Huang, Y.; Xing, L.; Stockwell, B.R.; et al. Lead compounds for the development of SARS-CoV-2 3CL protease inhibitors. *Nat Commun* **2021**, *12*, 2016, doi:10.1038/s41467-021-22362-2.
159. Martí, S.; Arafet, K.; Lodola, A.; Mulholland, A.J.; Świderek, K.; Moliner, V. Impact of Warhead Modulations on the Covalent Inhibition of SARS-CoV-2 M<sup>pro</sup> Explored by QM/MM Simulations. *ACS Catalysis* **2022**, 698-708, doi:10.1021/acscatal.1c04661.
160. Manandhar, A.; Srinivasulu, V.; Hamad, M.; Tarazi, H.; Omar, H.; Colussi, D.J.; Gordon, J.; Childers, W.; Klein, M.L.; Al-Tel, T.H.; et al. Discovery of Novel Small-Molecule Inhibitors of SARS-CoV-2 Main Protease as Potential Leads for COVID-19 Treatment. *J Chem Inf Model* **2021**, *61*, 4745-4757, doi:10.1021/acs.jcim.1c00684.
161. Zaidman, D.; Gehrtz, P.; Filep, M.; Fearon, D.; Gabizon, R.; Douangamath, A.; Prilusky, J.; Duberstein, S.; Cohen, G.; Owen, C.D.; et al. An automatic pipeline for the design of irreversible derivatives identifies a potent SARS-CoV-2 M<sup>pro</sup> inhibitor. *Cell Chem Biol* **2021**, *28*, 1795-1806.e1795, doi:10.1016/j.chembiol.2021.05.018.
162. Stille, J.K.; Tjutris, J.; Wang, G.; Venegas, F.A.; Hennecker, C.; Rueda, A.M.; Sharon, I.; Blaine, N.; Miron, C.E.; Pinus, S.; et al. Design, synthesis and in vitro evaluation of novel SARS-CoV-2 3CL<sup>pro</sup> covalent inhibitors. *Eur J Med Chem* **2021**, *229*, 114046, doi:10.1016/j.ejmech.2021.114046.
163. Gossen, J.; Albani, S.; Hanke, A.; Joseph, B.P.; Bergh, C.; Kuzikov, M.; Costanzi, E.; Manelfi, C.; Storici, P.; Gribbon, P.; et al. A Blueprint for High Affinity SARS-CoV-2 M<sup>pro</sup> Inhibitors from Activity-Based Compound Library Screening Guided by Analysis of Protein Dynamics. *ACS Pharmacol Transl Sci* **2021**, *4*, 1079-1095, doi:10.1021/acspstsci.0c00215.
164. Su, H.; Yao, S.; Zhao, W.; Zhang, Y.; Liu, J.; Shao, Q.; Wang, Q.; Li, M.; Xie, H.; Shang, W.; et al. Identification of pyrogallol as a warhead in design of covalent inhibitors for the SARS-CoV-2 3CL protease. *Nat Commun* **2021**, *12*, 3623, doi:10.1038/s41467-021-23751-3.
165. Xiao, T.; Cui, M.; Zheng, C.; Wang, M.; Sun, R.; Gao, D.; Bao, J.; Ren, S.; Yang, B.; Lin, J.; et al. Myricetin Inhibits SARS-CoV-2 Viral Replication by Targeting M<sup>pro</sup> and Ameliorates Pulmonary Inflammation. *Front Pharmacol* **2021**, *12*, 669642, doi:10.3389/fphar.2021.669642.
166. Chia, C.S.B. Novel Nitrile Peptidomimetics for Treating COVID-19. *ACS Med Chem Lett* **2022**, *13*, 330-331, doi:10.1021/acsmchemlett.2c00030.
167. Halford, B. The Path to Paxlovid. *ACS Central Science* **2022**, doi:10.1021/acscentsci.2c00369.
168. Owen, D.R.; Allerton, C.M.N.; Anderson, A.S.; Aschenbrenner, L.; Avery, M.; Berritt, S.; Boras, B.; Cardin, R.D.; Carlo, A.; Coffman, K.J.; et al. An oral SARS-CoV-2 M<sup>pro</sup> inhibitor clinical candidate for the treatment of COVID-19. *Science* **2021**, *374*, 1586-1593, doi:10.1126/science.abl4784.
169. COVID-19: EMA recommends conditional marketing authorisation for Paxlovid.

170. Lamb, Y.N. Nirmatrelvir Plus Ritonavir: First Approval. *Drugs* **2022**, doi:10.1007/s40265-022-01692-5.
171. Li, J.; Lin, C.; Zhou, X.; Zhong, F.; Zeng, P.; Yang, Y.; Zhang, Y.; Yu, B.; Fan, X.; McCormick, P.J.; et al. Structural Basis of the Main Proteases of Coronavirus Bound to Drug Candidate PF-07321332. *J Virol* **2022**, e0201321, doi:10.1128/jvi.02013-21.
172. Ramos-Guzmán, C.A.; Ruiz-Pernía, J.J.; Tuñón, I. Computational simulations on the binding and reactivity of a nitrile inhibitor of the SARS-CoV-2 main protease. *Chem Commun (Camb)* **2021**, 57, 9096-9099, doi:10.1039/d1cc03953a.
173. Ngo, S.T.; Nguyen, T.H.; Tung, N.T.; Mai, B.K. Insights into the binding and covalent inhibition mechanism of PF-07321332 to SARS-CoV-2 Mpro. *RSC Advances* **2022**, 12, 3729-3737, doi:10.1039/d1ra08752e.
174. Ullrich, S.; Ekanayake, K.B.; Otting, G.; Nitsche, C. Main protease mutants of SARS-CoV-2 variants remain susceptible to nirmatrelvir. *Bioorg Med Chem Lett* **2022**, 62, 128629, doi:10.1016/j.bmcl.2022.128629.
175. Abdelnabi, R.; Foo, C.S.; Jochmans, D.; Vangeel, L.; De Jonghe, S.; Augustijns, P.; Mols, R.; Weynand, B.; Wattanakul, T.; Hoglund, R.M.; et al. The oral protease inhibitor (PF-07321332) protects Syrian hamsters against infection with SARS-CoV-2 variants of concern. *Nat Commun* **2022**, 13, 719, doi:10.1038/s41467-022-28354-0.
176. Li, P.; Wang, Y.; Lavrijsen, M.; Lamers, M.M.; de Vries, A.C.; Rottier, R.J.; Bruno, M.J.; Peppelenbosch, M.P.; Haagmans, B.L.; Pan, Q. SARS-CoV-2 Omicron variant is highly sensitive to molnupiravir, nirmatrelvir, and the combination. *Cell Res* **2022**, 32, 322-324, doi:10.1038/s41422-022-00618-w.
177. Rai, D.K.; Yurgelonis, I.; McMonagle, P.; Rothan, H.A.; Hao, L.; Gribenko, A.; Titova, E.; Kreiswirth, B.; White, K.M.; Zhu, Y.; et al. Nirmatrelvir, an orally active Mpro inhibitor, is a potent inhibitor of SARS-CoV-2 Variants of Concern. *bioRxiv* **2022**, 2022.2001.2017.476644, doi:10.1101/2022.01.17.476644.
178. Greasley, S.E.; Noell, S.; Plotnikova, O.; Ferre, R.; Liu, W.; Bolanos, B.; Fennell, K.; Nicki, J.; Craig, T.; Zhu, Y.; et al. Structural basis for the in vitro efficacy of nirmatrelvir against SARS-CoV-2 variants. *J Biol Chem* **2022**, 101972, doi:10.1016/j.jbc.2022.101972.
179. Rosales, R.; McGovern, B.L.; Rodriguez, M.L.; Rai, D.K.; Cardin, R.D.; Anderson, A.S.; group, P.S.P.s.; Sordillo, E.M.; van Bakel, H.; Simon, V.; et al. Nirmatrelvir, Molnupiravir, and Remdesivir maintain potent *in vitro* activity against the SARS-CoV-2 Omicron variant. *bioRxiv* **2022**, 2022.2001.2017.476685, doi:10.1101/2022.01.17.476685.
180. Bai, B.; Arutyunova, E.; Khan, M.B.; Lu, J.; Joyce, M.A.; Saffran, H.A.; Shields, J.A.; Kandadai, A.S.; Belovodskiy, A.; Hena, M.; et al. Peptidomimetic nitrile warheads as SARS-CoV-2 3CL protease inhibitors. *RSC Med Chem* **2021**, 12, 1722-1730, doi:10.1039/d1md00247c.
181. Hattori, S.I.; Higshi-Kuwata, N.; Raghavaiah, J.; Das, D.; Bulut, H.; Davis, D.A.; Takamatsu, Y.; Matsuda, K.; Takamune, N.; Kishimoto, N.; et al. GRL-0920, an Indole Chloropyridinyl Ester, Completely Blocks SARS-CoV-2 Infection. *mBio* **2020**, 11, doi:10.1128/mBio.01833-20.
182. Ghosh, A.K.; Raghavaiah, J.; Shahabi, D.; Yadav, M.; Anson, B.J.; Lendy, E.K.; Hattori, S.I.; Higashi-Kuwata, N.; Mitsuya, H.; Mesecar, A.D. Indole Chloropyridinyl Ester-Derived SARS-CoV-2 3CLpro Inhibitors: Enzyme Inhibition, Antiviral Efficacy, Structure-Activity Relationship, and X-ray Structural Studies. *J Med Chem* **2021**, 64, 14702-14714, doi:10.1021/acs.jmedchem.1c01214.

183. Breidenbach, J.; Lemke, C.; Pillaiyar, T.; Schäkel, L.; Al Hamwi, G.; Dieltz, M.; Gedschold, R.; Geiger, N.; Lopez, V.; Mirza, S.; et al. Targeting the Main Protease of SARS-CoV-2: From the Establishment of High Throughput Screening to the Design of Tailored Inhibitors. *Angew Chem Int Ed Engl* **2021**, *60*, 10423-10429, doi:10.1002/anie.202016961.
184. Ghosh, A.K.; Shahabi, D.; Yadav, M.; Kovala, S.; Anson, B.J.; Lendy, E.K.; Bonham, C.; Sirohi, D.; Brito-Sierra, C.A.; Hattori, S.I.; et al. Chloropyridinyl Esters of Nonsteroidal Anti-Inflammatory Agents and Related Derivatives as Potent SARS-CoV-2 3CL Protease Inhibitors. *Molecules* **2021**, *26*, doi:10.3390/molecules26195782.
185. Nakamura, Y.; Feng, Q.; Kumagai, T.; Torikai, K.; Ohigashi, H.; Osawa, T.; Noguchi, N.; Niki, E.; Uchida, K. Ebselen, a glutathione peroxidase mimetic seleno-organic compound, as a multifunctional antioxidant. Implication for inflammation-associated carcinogenesis. *J Biol Chem* **2002**, *277*, 2687-2694, doi:10.1074/jbc.M109641200.
186. Sakurai, T.; Kanayama, M.; Shibata, T.; Itoh, K.; Kobayashi, A.; Yamamoto, M.; Uchida, K. Ebselen, a seleno-organic antioxidant, as an electrophile. *Chem Res Toxicol* **2006**, *19*, 1196-1204, doi:10.1021/tx0601105.
187. Monajemi, H.; Zain, S.M. Strong inhibition of M-Protease activity of Coronavirus by using PX-12 inhibitor based on ab initio ONIOM calculations. *Journal of Chemical Research* **2021**, *45*, 136-140, doi:10.1177/1747519820938025.
188. Madabeni, A.; Nogara, P.A.; Omege, F.B.; Rocha, J.B.T.; Orian, L. Mechanistic insight into sars-cov-2 mpro inhibition by organoselenides: The ebselen case study. *Applied Sciences (Switzerland)* **2021**, *11*, doi:10.3390/app11146291.
189. Parise, A.; Romeo, I.; Russo, N.; Marino, T. The Se-S Bond Formation in the Covalent Inhibition Mechanism of SARS-CoV-2 Main Protease by Ebselen-like Inhibitors: A Computational Study. *Int J Mol Sci* **2021**, *22*, doi:10.3390/ijms22189792.
190. Amporndanai, K.; Meng, X.; Shang, W.; Jin, Z.; Rogers, M.; Zhao, Y.; Rao, Z.; Liu, Z.J.; Yang, H.; Zhang, L.; et al. Inhibition mechanism of SARS-CoV-2 main protease by ebselen and its derivatives. *Nat Commun* **2021**, *12*, 3061, doi:10.1038/s41467-021-23313-7.
191. Menéndez, C.A.; Byléhn, F.; Perez-Lemus, G.R.; Alvarado, W.; de Pablo, J.J. Molecular characterization of ebselen binding activity to SARS-CoV-2 main protease. *Sci Adv* **2020**, *6*, doi:10.1126/sciadv.abd0345.
192. Zmudzinski, M.; Rut, W.; Olech, K.; Granda, J.; Giurg, M.; Burda-Grabowska, M.; Zhang, L.; Sun, X.; Lv, Z.; Nayak, D.; et al. Ebselen derivatives are very potent dual inhibitors of SARS-CoV-2 proteases - PL<sup>pro</sup> and M<sup>pro</sup> in in vitro studies. *bioRxiv* **2020**, 2020.2008.2030.273979, doi:10.1101/2020.08.30.273979.
193. Huff, S.; Kummetha, I.R.; Tiwari, S.K.; Huante, M.B.; Clark, A.E.; Wang, S.; Bray, W.; Smith, D.; Carlin, A.F.; Endsley, M.; et al. Discovery and Mechanism of SARS-CoV-2 Main Protease Inhibitors. *J Med Chem* **2021**, doi:10.1021/acs.jmedchem.1c00566.
194. Sun, L.Y.; Chen, C.; Su, J.; Li, J.Q.; Jiang, Z.; Gao, H.; Chigan, J.Z.; Ding, H.H.; Zhai, L.; Yang, K.W. Ebsulfur and Ebselen as highly potent scaffolds for the development of potential SARS-CoV-2 antivirals. *Bioorg Chem* **2021**, *112*, 104889, doi:10.1016/j.bioorg.2021.104889.
195. Chen, W.; Feng, B.; Han, S.; Wang, P.; Zang, Y.; Li, J.; Hu, Y. Discovery of highly potent SARS-CoV-2 M<sup>pro</sup> inhibitors based on benzoisothiazolone scaffold. *Bioorg Med Chem Lett* **2022**, *58*, 128526, doi:10.1016/j.bmcl.2022.128526.



196. Ghosh, A.K.; Gong, G.; Grum-Tokars, V.; Mulhearn, D.C.; Baker, S.C.; Coughlin, M.; Prabhakar, B.S.; Sleeman, K.; Johnson, M.E.; Mesecar, A.D. Design, synthesis and antiviral efficacy of a series of potent chloropyridyl ester-derived SARS-CoV 3CLpro inhibitors. *Bioorg Med Chem Lett* **2008**, *18*, 5684-5688, doi:10.1016/j.bmcl.2008.08.082.
197. Kirsch, P.; Hartman, A.M.; Hirsch, A.K.H.; Empting, M. Concepts and Core Principles of Fragment-Based Drug Design. *Molecules* **2019**, *24*, doi:10.3390/molecules24234309.
198. McAulay, K.; Bilsland, A.; Bon, M. Reactivity of Covalent Fragments and Their Role in Fragment Based Drug Discovery. *Pharmaceuticals (Basel)* **2022**, *15*, doi:10.3390/ph15111366.
199. Hacker, S.M.; Backus, K.M.; Lazear, M.R.; Forli, S.; Correia, B.E.; Cravatt, B.F. Global profiling of lysine reactivity and ligandability in the human proteome. *Nat Chem* **2017**, *9*, 1181-1190, doi:10.1038/nchem.2826.
200. Wamser, R.; Pach, S.; Arkona, C.; Baumgardt, M.; Aziz, U.B.A.; Hocke, A.C.; Wolber, G.; Rademann, J. A Critical Study on Acylating and Covalent Reversible Fragment Inhibitors of SARS-CoV-2 Main Protease Targeting the S1 Site with Pyridine. *ChemMedChem* **2023**, *18*, e202200635, doi:10.1002/cmdc.202200635.
201. Dulsat, J.; López-Nieto, B.; Estrada-Tejedor, R.; Borrell, J.I. Evaluation of Free Online ADMET Tools for Academic or Small Biotech Environments. *Molecules* **2023**, *28*, doi:10.3390/molecules28020776.
202. Zhu, K.; Borrelli, K.W.; Greenwood, J.R.; Day, T.; Abel, R.; Farid, R.S.; Harder, E. Docking covalent inhibitors: a parameter free approach to pose prediction and scoring. *J Chem Inf Model* **2014**, *54*, 1932-1940, doi:10.1021/ci500118s.
203. Toledo Warshaviak, D.; Golan, G.; Borrelli, K.W.; Zhu, K.; Kalid, O. Structure-based virtual screening approach for discovery of covalently bound ligands. *J Chem Inf Model* **2014**, *54*, 1941-1950, doi:10.1021/ci500175r.
204. Schrödinger Release 2017-2, LigPrep ; Schrödinger, LLC: New York, NY, USA, 2017.
205. Schrödinger Release 2017-2, Schrödinger Suite 2017-2 Protein Preparation Wizard ; Epik, Schrödinger, LLC: New York, NY, USA, 2017.
206. Banks, J.L.; Beard, H.S.; Cao, Y.; Cho, A.E.; Damm, W.; Farid, R.; Felts, A.K.; Halgren, T.A.; Mainz, D.T.; Maple, J.R.; et al. Integrated Modeling Program, Applied Chemical Theory (IMPACT). *J Comput Chem* **2005**, *26*, 1752-1780, doi:10.1002/jcc.20292.
207. Berman, H.M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.N.; Weissig, H.; Shindyalov, I.N.; Bourne, P.E. The Protein Data Bank. *Nucleic Acids Res* **2000**, *28*, 235-242, doi:10.1093/nar/28.1.235.
208. RCSB PDB. Available online: [www.rcsb.org](http://www.rcsb.org) (accessed on 29 June 2022).
209. Sastry, G.M.; Adzhigirey, M.; Day, T.; Annabhimoju, R.; Sherman, W. Protein and ligand preparation: parameters, protocols, and influence on virtual screening enrichments. *J Comput Aided Mol Des* **2013**, *27*, 221-234, doi:10.1007/s10822-013-9644-8.
210. Sherman, W.; Beard, H.S.; Farid, R. Use of an induced fit receptor structure in virtual screening. *Chem Biol Drug Des* **2006**, *67*, 83-84, doi:10.1111/j.1747-0285.2005.00327.x.
211. Sherman, W.; Day, T.; Jacobson, M.P.; Friesner, R.A.; Farid, R. Novel procedure for modeling ligand/receptor induced fit effects. *J Med Chem* **2006**, *49*, 534-553, doi:10.1021/jm050540c.

212. Zhou, D.; Liu, F.; Zheng, Y.; Hu, L.; Huang, T.; Huang, Y.S. Deffini: A family-specific deep neural network model for structure-based virtual screening. *Comput Biol Med* **2022**, *151*, 106323, doi:10.1016/j.combiomed.2022.106323.
213. Grimm, M.; Liu, Y.; Yang, X.; Bu, C.; Xiao, Z.; Cao, Y. LigMate: A Multifeature Integration Algorithm for Ligand-Similarity-Based Virtual Screening. *J Chem Inf Model* **2020**, *60*, 6044-6053, doi:10.1021/acs.jcim.9b01210.
214. Stepišnik, T.; Škrlić, B.; Wicker, J.; Kocev, D. A comprehensive comparison of molecular feature representations for use in predictive modeling. *Comput Biol Med* **2021**, *130*, 104197, doi:10.1016/j.combiomed.2020.104197.
215. Tiikkainen, P.; Markt, P.; Wolber, G.; Kirchmair, J.; Distinto, S.; Poso, A.; Kallioniemi, O. Critical comparison of virtual screening methods against the MUV data set. *J Chem Inf Model* **2009**, *49*, 2168-2178, doi:10.1021/ci900249b.
216. Altae-Tran, H.; Ramsundar, B.; Pappu, A.S.; Pande, V. Low Data Drug Discovery with One-Shot Learning. *ACS Cent Sci* **2017**, *3*, 283-293, doi:10.1021/acscentsci.6b00367.
217. Winter, R.; Montanari, F.; Noé, F.; Clevert, D.A. Learning continuous and data-driven molecular descriptors by translating equivalent chemical representations. *Chem Sci* **2019**, *10*, 1692-1701, doi:10.1039/c8sc04175j.
218. Kearnes, S.; McCloskey, K.; Berndl, M.; Pande, V.; Riley, P. Molecular graph convolutions: moving beyond fingerprints. *J Comput Aided Mol Des* **2016**, *30*, 595-608, doi:10.1007/s10822-016-9938-8.
219. Li, P.; Li, Y.; Hsieh, C.Y.; Zhang, S.; Liu, X.; Liu, H.; Song, S.; Yao, X. TrimNet: learning molecular representation from triplet messages for biomedicine. *Brief Bioinform* **2021**, *22*, doi:10.1093/bib/bbaa266.
220. Armstrong, M.S.; Morris, G.M.; Finn, P.W.; Sharma, R.; Moretti, L.; Cooper, R.I.; Richards, W.G. ElectroShape: fast molecular similarity calculations incorporating shape, chirality and electrostatics. *J Comput Aided Mol Des* **2010**, *24*, 789-801, doi:10.1007/s10822-010-9374-0.
221. Rohrer, S.G.; Baumann, K. Maximum unbiased validation (MUV) data sets for virtual screening based on PubChem bioactivity data. *J Chem Inf Model* **2009**, *49*, 169-184, doi:10.1021/ci8002649.
222. Tran-Nguyen, V.K.; Jacquemard, C.; Rognan, D. LIT-PCBA: An Unbiased Data Set for Machine Learning and Virtual Screening. *J Chem Inf Model* **2020**, *60*, 4263-4273, doi:10.1021/acs.jcim.0c00155.
223. Duch, W.; Swaminathan, K.; Meller, J. Artificial intelligence approaches for rational drug design and discovery. *Curr Pharm Des* **2007**, *13*, 1497-1508, doi:10.2174/138161207780765954.
224. Crampon, K.; Giorkallos, A.; Deldossi, M.; Baud, S.; Steffanel, L.A. Machine-learning methods for ligand-protein molecular docking. *Drug Discov Today* **2022**, *27*, 151-164, doi:10.1016/j.drudis.2021.09.007.
225. Rifaioglu, A.S.; Atas, H.; Martin, M.J.; Cetin-Atalay, R.; Atalay, V.; Doğan, T. Recent applications of deep learning and machine intelligence on in silico drug discovery: methods, tools and databases. *Brief Bioinform* **2019**, *20*, 1878-1912, doi:10.1093/bib/bby061.
226. Plewczynski, D.; Spieser, S.A.; Koch, U. Performance of machine learning methods for ligand-based virtual screening. *Comb Chem High Throughput Screen* **2009**, *12*, 358-368, doi:10.2174/138620709788167962.

227. Warren, G.L.; Andrews, C.W.; Capelli, A.M.; Clarke, B.; LaLonde, J.; Lambert, M.H.; Lindvall, M.; Nevins, N.; Semus, S.F.; Senger, S.; et al. A critical assessment of docking programs and scoring functions. *J Med Chem* **2006**, *49*, 5912-5931, doi:10.1021/jm050362n.
228. Braga, R.C.; Andrade, C.H. Assessing the performance of 3D pharmacophore models in virtual screening: how good are they? *Curr Top Med Chem* **2013**, *13*, 1127-1138, doi:10.2174/1568026611313090010.
229. Zhou, Z.; Felts, A.K.; Friesner, R.A.; Levy, R.M. Comparative performance of several flexible docking programs and scoring functions: enrichment studies for a diverse set of pharmaceutically relevant targets. *J Chem Inf Model* **2007**, *47*, 1599-1608, doi:10.1021/ci7000346.
230. Chen, Z.; Li, H.L.; Zhang, Q.J.; Bao, X.G.; Yu, K.Q.; Luo, X.M.; Zhu, W.L.; Jiang, H.L. Pharmacophore-based virtual screening versus docking-based virtual screening: a benchmark comparison against eight targets. *Acta Pharmacol Sin* **2009**, *30*, 1694-1708, doi:10.1038/aps.2009.159.
231. Huang, S.Y.; Grinter, S.Z.; 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.
232. Ashtawy, H.M.; Mahapatra, N.R. A comparative assessment of ranking accuracies of conventional and machine-learning-based scoring functions for protein-ligand binding affinity prediction. *IEEE/ACM Trans Comput Biol Bioinform* **2012**, *9*, 1301-1313, doi:10.1109/TCBB.2012.36.
233. Wang, R.; Lu, Y.; Fang, X.; Wang, S. An extensive test of 14 scoring functions using the PDBbind refined set of 800 protein-ligand complexes. *J Chem Inf Comput Sci* **2004**, *44*, 2114-2125, doi:10.1021/ci049733j.
234. Ferrara, P.; Gohlke, H.; Price, D.J.; Klebe, G.; Brooks, C.L. Assessing scoring functions for protein-ligand interactions. *J Med Chem* **2004**, *47*, 3032-3047, doi:10.1021/jm030489h.
235. Cheng, T.; Li, X.; Li, Y.; Liu, Z.; Wang, R. Comparative assessment of scoring functions on a diverse test set. *J Chem Inf Model* **2009**, *49*, 1079-1093, doi:10.1021/ci9000053.
236. Huang, N.; Shoichet, B.K.; Irwin, J.J. Benchmarking sets for molecular docking. *J Med Chem* **2006**, *49*, 6789-6801, doi:10.1021/jm0608356.
237. Mysinger, M.M.; Carchia, M.; Irwin, J.J.; Shoichet, B.K. Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking. *J Med Chem* **2012**, *55*, 6582-6594, doi:10.1021/jm300687e.
238. Wallach, I.; Lilien, R. Virtual decoy sets for molecular docking benchmarks. *J Chem Inf Model* **2011**, *51*, 196-202, doi:10.1021/ci100374f.
239. Gatica, E.A.; Cavasotto, C.N. Ligand and decoy sets for docking to G protein-coupled receptors. *J Chem Inf Model* **2012**, *52*, 1-6, doi:10.1021/ci200412p.
240. Vogel, S.M.; Bauer, M.R.; Boeckler, F.M. DEKOIS: demanding evaluation kits for objective in silico screening--a versatile tool for benchmarking docking programs and scoring functions. *J Chem Inf Model* **2011**, *51*, 2650-2665, doi:10.1021/ci2001549.
241. Bauer, M.R.; Ibrahim, T.M.; Vogel, S.M.; Boeckler, F.M. Evaluation and optimization of virtual screening workflows with DEKOIS 2.0--a public library of challenging docking benchmark sets. *J Chem Inf Model* **2013**, *53*, 1447-1462, doi:10.1021/ci400115b.
242. Jahn, A.; Hinselmann, G.; Fechner, N.; Zell, A. Optimal assignment methods for ligand-based virtual screening. *J Cheminform* **2009**, *1*, 14, doi:10.1186/1758-2946-1-14.

243. Ripphausen, P.; Wassermann, A.M.; Bajorath, J. REPROVIS-DB: a benchmark system for ligand-based virtual screening derived from reproducible prospective applications. *J Chem Inf Model* **2011**, *51*, 2467-2473, doi:10.1021/ci200309j.
244. Ugurlu, S.Y.; McDonald, D.; Lei, H.; Jones, A.M.; Li, S.; Tong, H.Y.; Butler, M.S.; He, S. Cobdock: an accurate and practical machine learning-based consensus blind docking method. *J Cheminform* **2024**, *16*, 5, doi:10.1186/s13321-023-00793-x.
245. Zhou, H.; Skolnick, J. FRAGSITE2: A structure and fragment-based approach for virtual ligand screening. *Protein Sci* **2024**, *33*, e4869, doi:10.1002/pro.4869.
246. Morris, C.J.; Stern, J.A.; Stark, B.; Christopherson, M.; Della Corte, D. MILCDock: Machine Learning Enhanced Consensus Docking for Virtual Screening in Drug Discovery. *J Chem Inf Model* **2022**, *62*, 5342-5350, doi:10.1021/acs.jcim.2c00705.
247. Zhang, X.; Shen, C.; Jiang, D.; Zhang, J.; Ye, Q.; Xu, L.; Hou, T.; Pan, P.; Kang, Y. TB-IECS: an accurate machine learning-based scoring function for virtual screening. *J Cheminform* **2023**, *15*, 63, doi:10.1186/s13321-023-00731-x.
248. Svensson, E.; Hoedt, P.J.; Hochreiter, S.; Klambauer, G. HyperPCM: Robust Task-Conditioned Modeling of Drug-Target Interactions. *J Chem Inf Model* **2024**, *64*, 2539-2553, doi:10.1021/acs.jcim.3c01417.
249. Neves, M.A.; Totrov, M.; Abagyan, R. Docking and scoring with ICM: the benchmarking results and strategies for improvement. *J Comput Aided Mol Des* **2012**, *26*, 675-686, doi:10.1007/s10822-012-9547-0.
250. Brozell, S.R.; Mukherjee, S.; Balius, T.E.; Roe, D.R.; Case, D.A.; Rizzo, R.C. Evaluation of DOCK 6 as a pose generation and database enrichment tool. *J Comput Aided Mol Des* **2012**, *26*, 749-773, doi:10.1007/s10822-012-9565-y.
251. Fan, H.; Irwin, J.J.; Webb, B.M.; Klebe, G.; Shoichet, B.K.; Sali, A. Molecular docking screens using comparative models of proteins. *J Chem Inf Model* **2009**, *49*, 2512-2527, doi:10.1021/ci9003706.
252. Good, A.C.; Oprea, T.I. Optimization of CAMD techniques 3. Virtual screening enrichment studies: a help or hindrance in tool selection? *J Comput Aided Mol Des* **2008**, *22*, 169-178, doi:10.1007/s10822-007-9167-2.
253. Wang, Y.; Xiao, J.; Suzek, T.O.; Zhang, J.; Wang, J.; Zhou, Z.; Han, L.; Karapetyan, K.; Dracheva, S.; Shoemaker, B.A.; et al. PubChem's BioAssay Database. *Nucleic Acids Res* **2012**, *40*, D400-412, doi:10.1093/nar/gkr1132.
254. Wang, H.; Liang, Q.; Hancock, J.T.; Khoshgoftaar, T.M. Feature selection strategies: a comparative analysis of SHAP-value and importance-based methods.
255. Md Shamim Hussain, M.J.Z., Dharmashankar Subramanian. Triplet Interaction Improves Graph Transformers: Accurate Molecular Graph Learning with Triplet Graph Transformers. **2024**.
256. Yuquan Li, C.-Y.H., Ruiqiang Lu , Xiaoqing Gong , Xiaorui Wang , Pengyong li , Shuo Liu , Yanan Tian , Dejun Jiang, Jiaxian Yan , Qifeng Bai , Huanxiang Liu , Shengyu Zhang , Xiaojun Yao. GLAM: An adaptive graph learning method for automated molecular interactions and properties predictions. **2022**.
257. Chen, L.; Tan, X.; Wang, D.; Zhong, F.; Liu, X.; Yang, T.; Luo, X.; Chen, K.; Jiang, H.; Zheng, M. TransformerCPI: improving compound-protein interaction prediction by sequence-based deep learning with self-attention mechanism and label reversal experiments. *Bioinformatics* **2020**, *36*, 4406-4414, doi:10.1093/bioinformatics/btaa524.

258. Jiang, M.; Li, Z.; Zhang, S.; Wang, S.; Wang, X.; Yuan, Q.; Wei, Z. Drug-target affinity prediction using graph neural network and contact maps. *RSC Adv* **2020**, *10*, 20701-20712, doi:10.1039/d0ra02297g.

## List of Abbreviations

**1%EF** – 1% Enrichment Factor

**3'UTR** – 3' Untranslated Region

**3CL<sup>PRO</sup>** – 3C-Like Protease

**4DES** – 4-Dimensional ElectroShape

**5'UTR** – 5' Untranslated Region

**5DES** – 5-Dimensional ElectroShape

**β-CoV** – β-CoronaVirus

**ACE2** – Angiotensin-Converting Enzyme 2

**ADME** – Absorption, Distribution, Metabolism, and Excretion

**AI** – Artificial Intelligence

**APP** – AntiProliferative Predictor

**AUC** – Area Under the Curve

**BPT** – Biotarget Predictor Tool

**CDKs** – Cyclin-Dependent Kinases

**CoVDock** – Covalent Docking

**COVID-19** – Coronavirus Disease-19

**CSR** – Chiral Shape Recognition

**DAS** – Drudit Affinity Score

**DEKOIS** – Demanding Evaluation Kits for Objective *in Silico* Screening

**DRUDIT** – DRUG Discovery Tools

**DUD** – Directory of Useful Decoys

**DUD-E** – DUD-Enhanced

**E** – Envelope

**EGFR** – Epidermal Growth Factor Receptor

**ES** – ElectroShape

**EUA** – Emergency Use Authorization

**FBDD** – Fragment-Based Drug Discovery

**FCoV** – Feline CoronaVirus

**FDA** – Food and Drug Administration

**FPR** – False Positive Rate

**GB** – Gradient Boosting

**GDD** – GPCRs Decoy Database

**GLL** – GPCRs Ligand Library

**GPCRs** – G Protein-Coupled Receptors

**GRAS** – Generally Recognized As Safe

**HBAs** – Hydrogen Bond Acceptors

**HBDs** – Hydrogen Bond Donors

**HTS** – High-Throughput Screening

**IFD** – Induced Fit Docking

**ISG15** – Interferon-Stimulated Gene 15

**LIT-PCBA** – LIT-PubChemBioActivity

**M** – Membrane

**MD** – Molecular Descriptors

**MERS-CoV** – Middle East Respiratory Syndrome Coronavirus

**ML** – Machine Learning

**MOLDESTO** – MOlecular DEScriptor TOol

**M<sup>PRO</sup>** – Main Protease

**MUV** – Maximun Unbiased Validation

**MW** – Molecular Weight

**N** – Nucleocapside

**NADH** – Nicotinamide Adenine Dinucleotide

**NN** – Neural Networks

**NRLiSt BDB** – Nuclear Receptors Ligands and Structures Benchmarking DataBase

**NSPs** – Non-Structural Proteins

**OPLS** – Optimized Potentials for Liquid Simulations

**ORF** – Open Reading Frame

**PCBioAssay** – Primary and Confirmatory BioAssays

**PL<sup>PRO</sup>** – Papain-Like Protease

**QSAR** – Quantitative Structure-Activity Relationships

**RBD** – Receptor Binding Domain

**RBs** – Rotatable Bonds

**REPROVIS** – REPROducible VIRTual Screens

**RF** – Random Forest

**ROC-AUC** – Receiver Operating Characteristic - Area Under the Curve

**RTC** – Replication-Transcription Complex

**S** – Spike

**SAR** – Structure-Activity Relationship

**SARS-CoV-2** – Severe Acute Respiratory Syndrome CoronaVirus 2

**SBVS** – Structure Based Virtual Screening

**SHAP** – SHapley Additive exPlanations

**ssRNA+** – single-stranded RNA +

**SVM** – Support Vector Machine

**Tc** – Tanimoto Coefficient

**TMPRSS2** – TransMembrane PRoteaSe, Serine 2

**TPR** – True Positive Rate

**TS** – Transition State

**USR** – Ultrafast Shape Recognition

**WHO** – World Health Organization

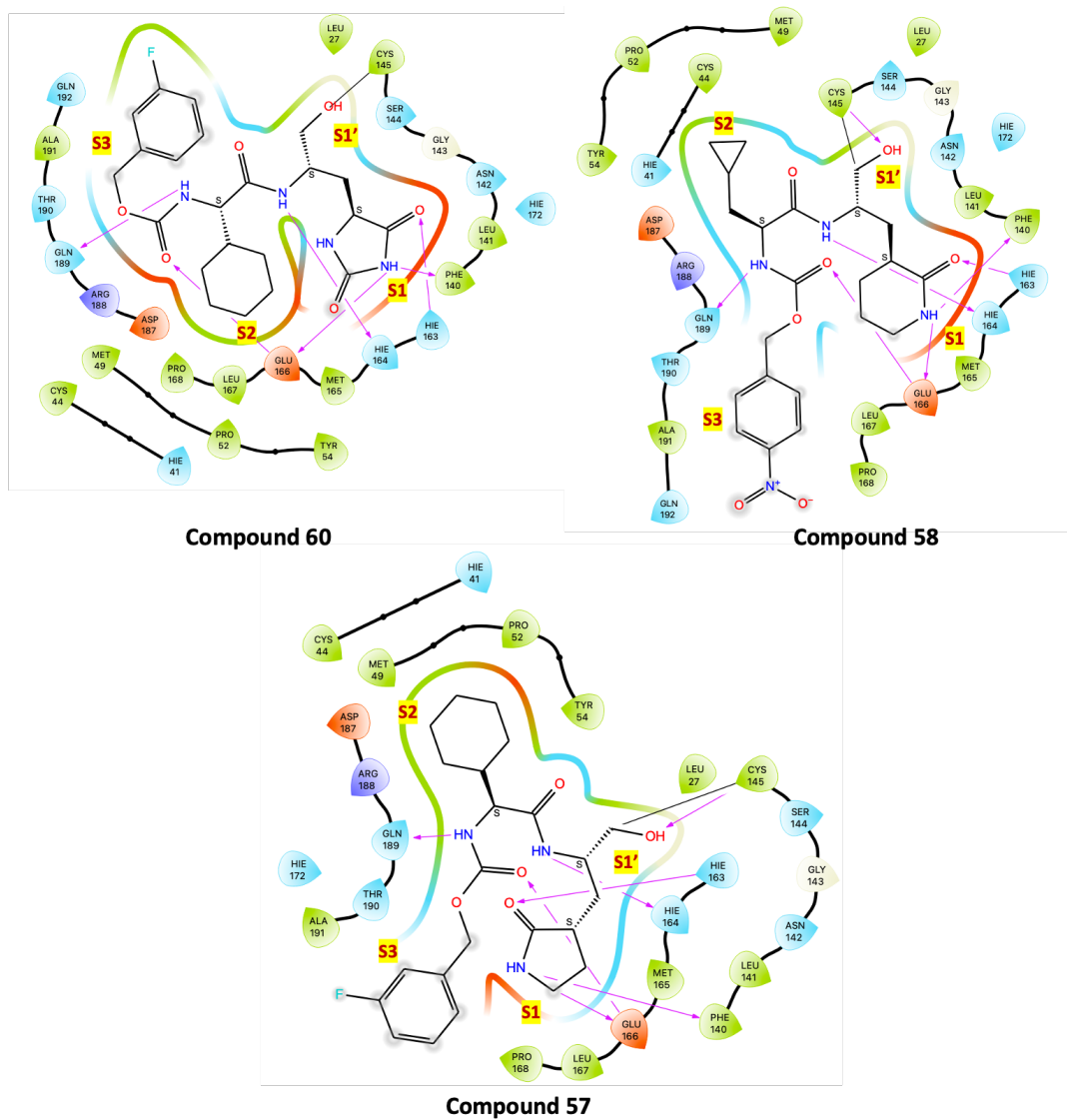
**VDS** – Virtual Decoy Sets



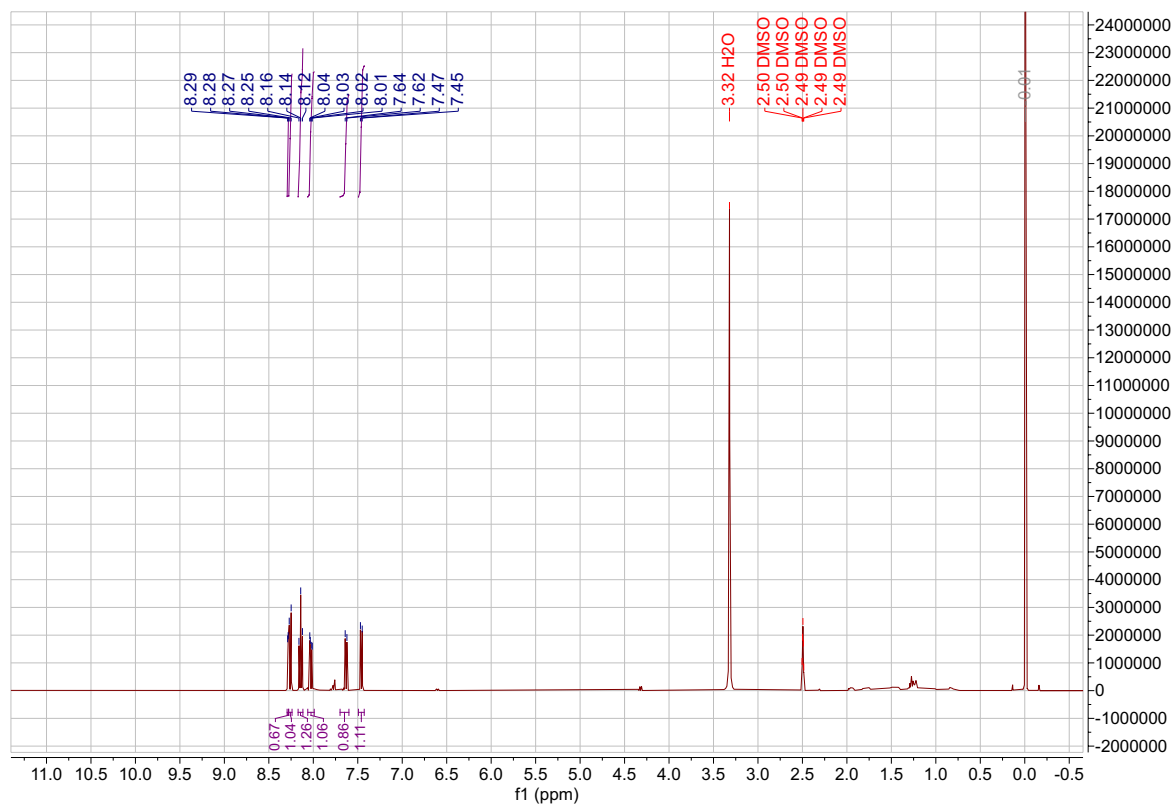
## Supplementary Material

The following are available online at: [Supporting Information - PhD Thesis Alessia Bono](#)

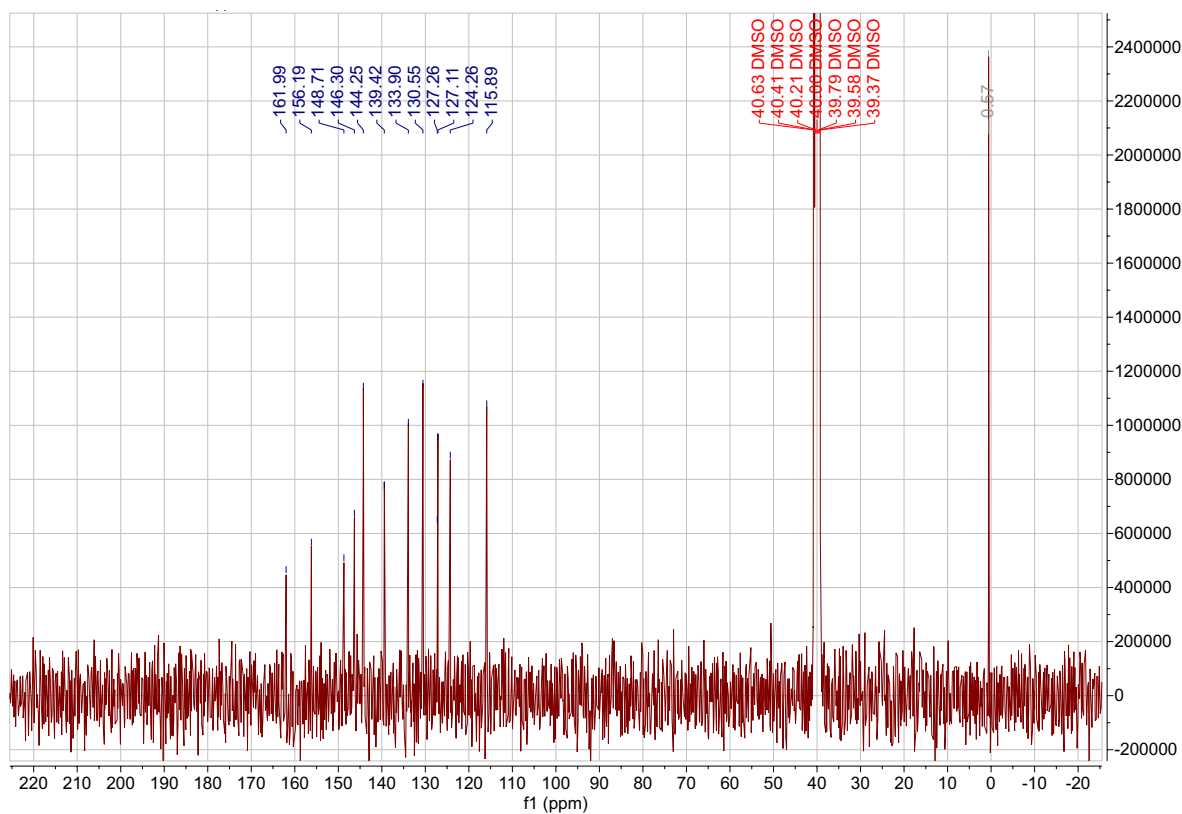
## Supplementary Figures



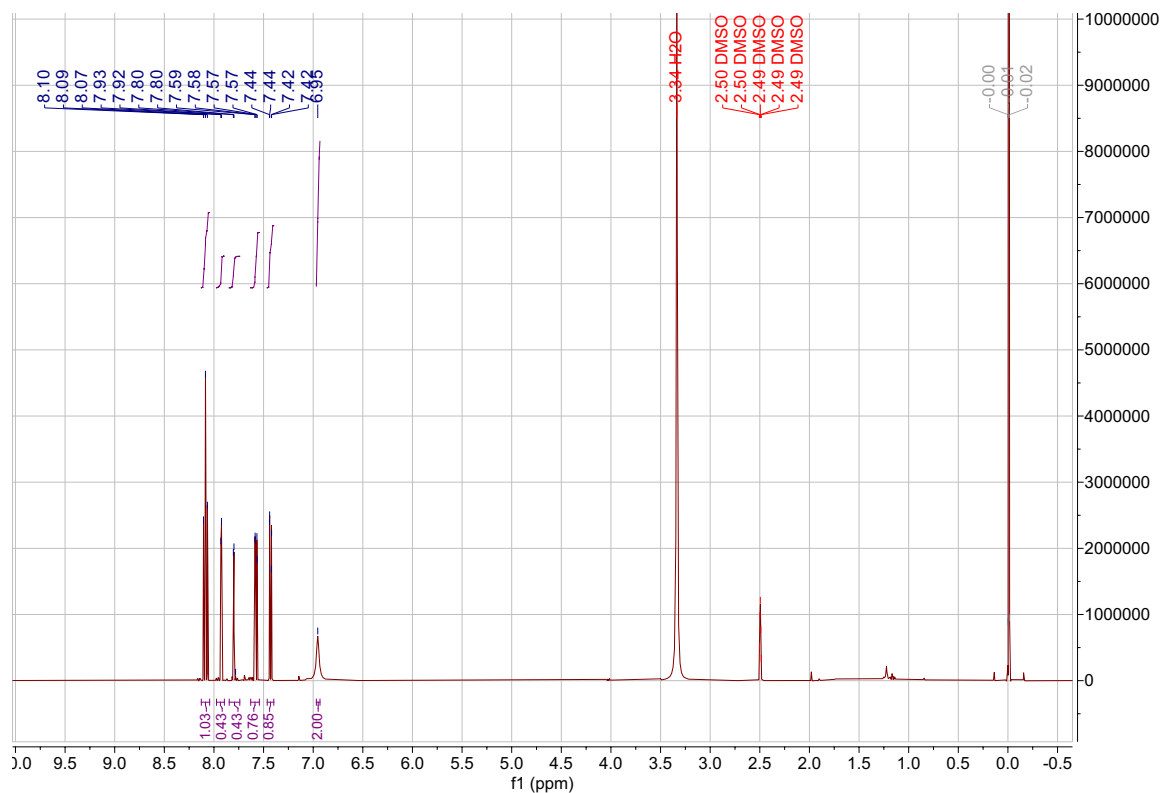
**Supplementary Figure S1.** 2D ligand interaction maps for compounds 57, 58, and 60.



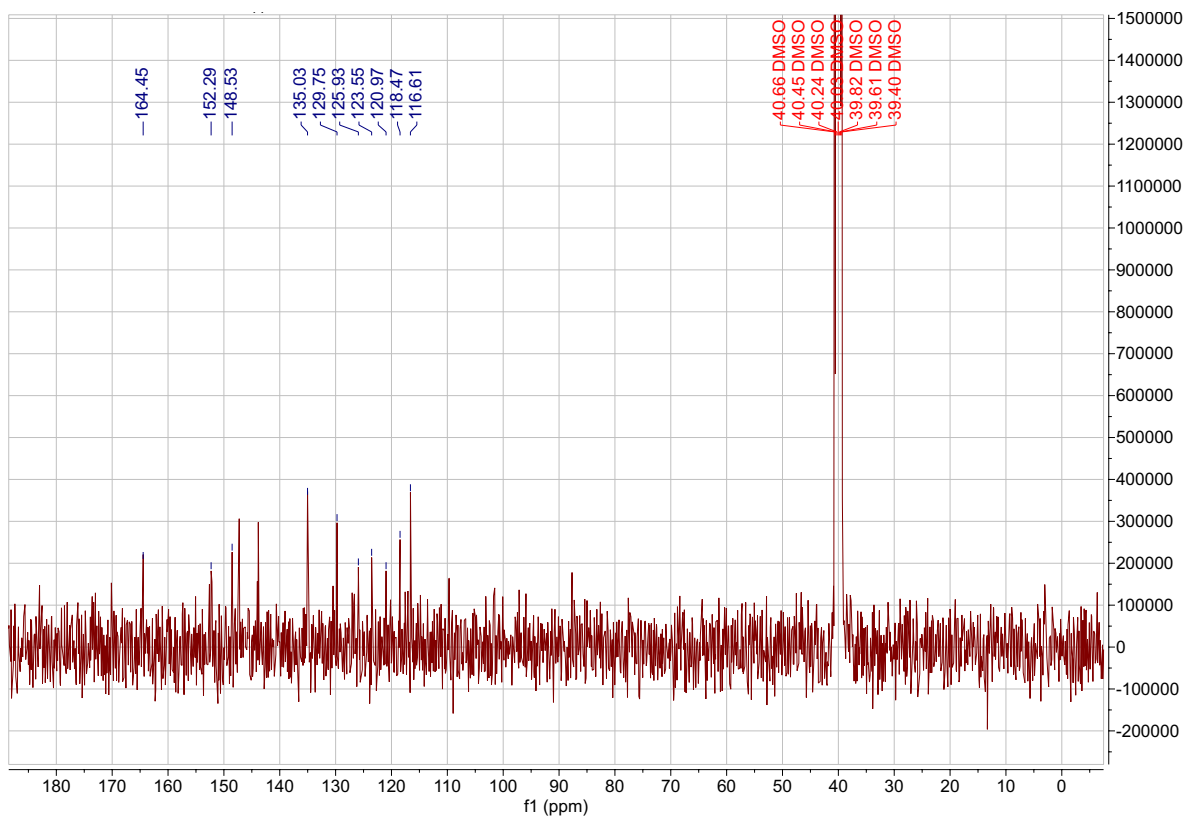
Supplementary Figure S2. <sup>1</sup>H NMR of compound 45a (DMSO-d<sub>6</sub>).



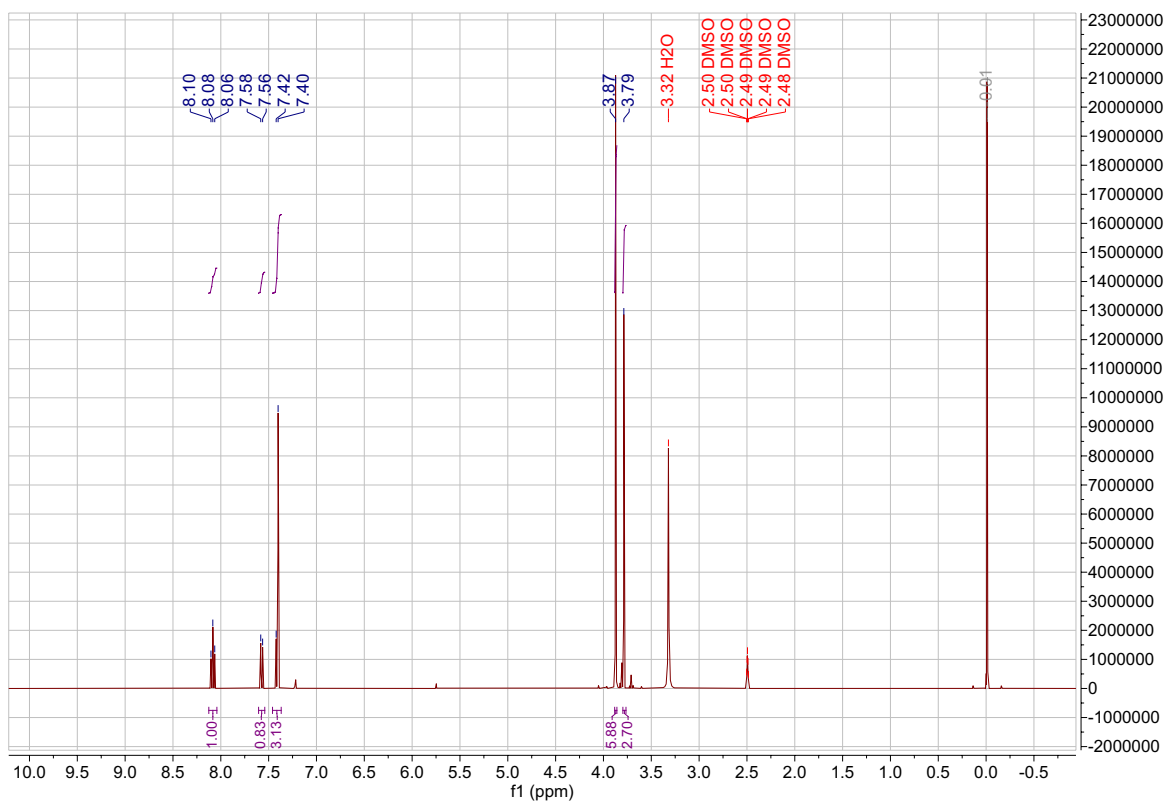
Supplementary Figure S3. <sup>13</sup>C NMR of compound 45a (DMSO-d<sub>6</sub>).



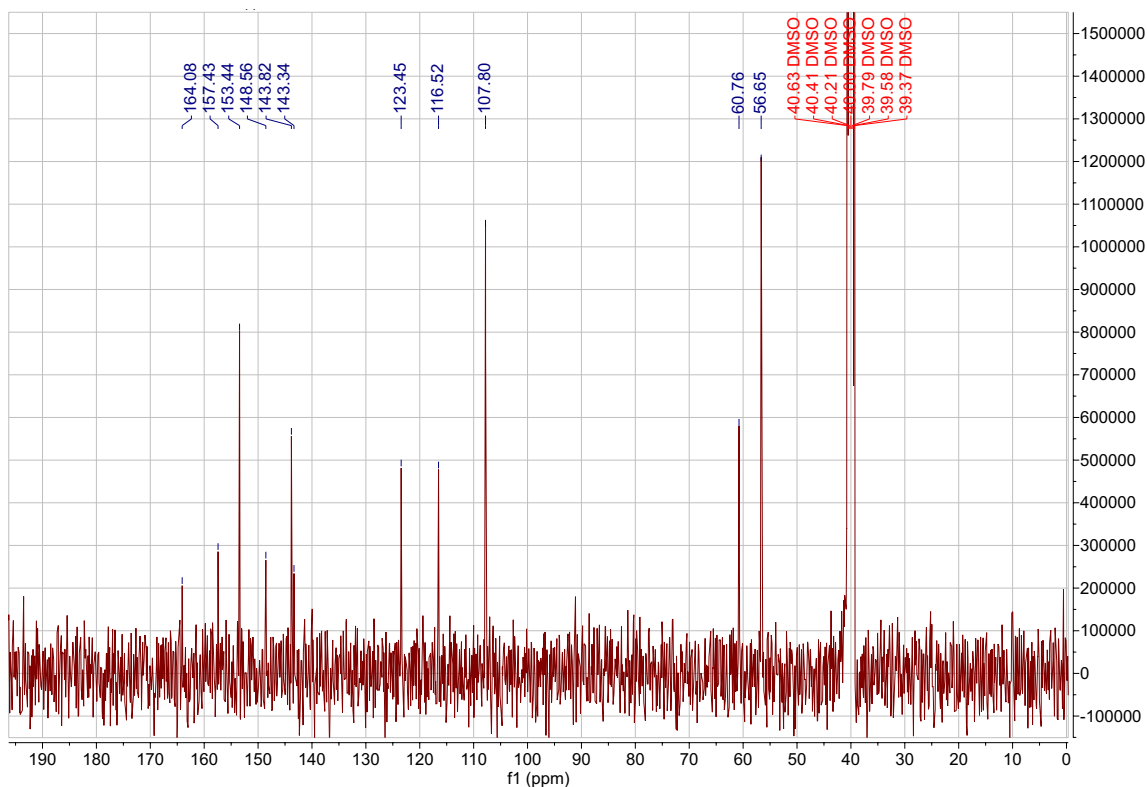
**Supplementary Figure S4: <sup>1</sup>H NMR compound 45b (DMSO-d<sub>6</sub>).**



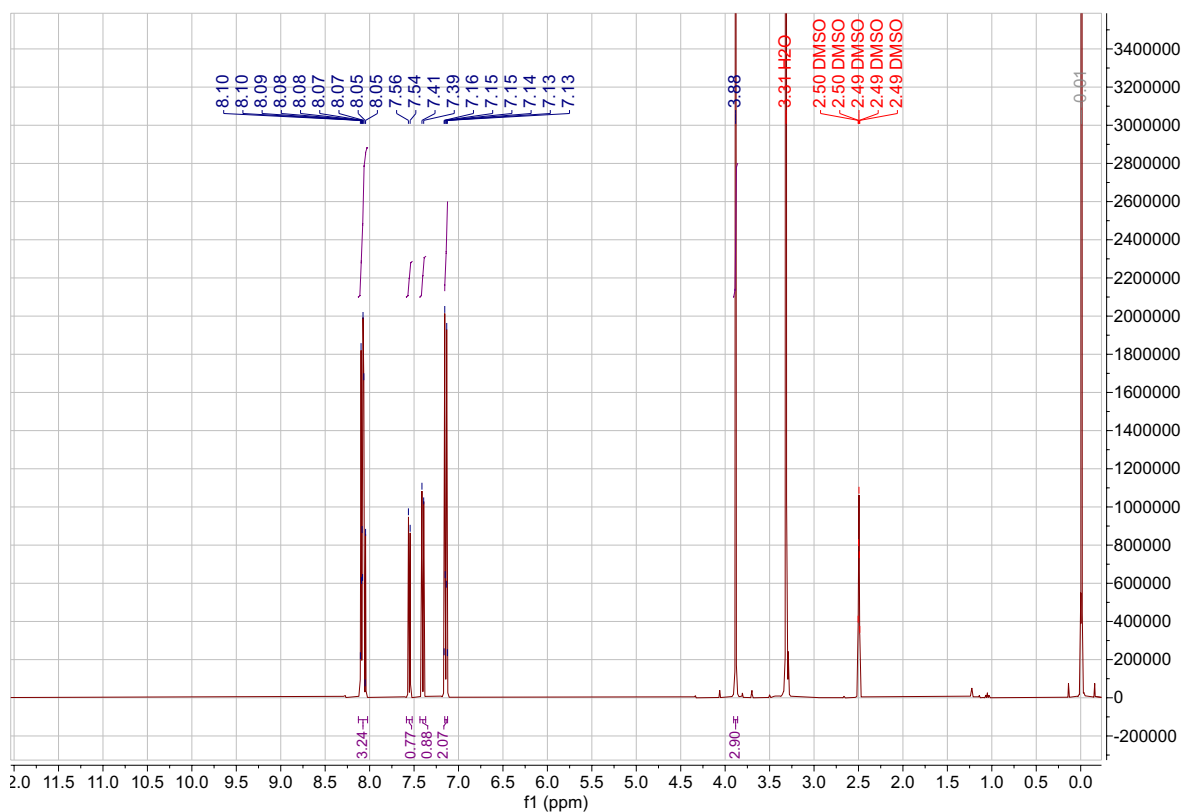
**Supplementary Figure S5: <sup>13</sup>C NMR compound 45b (DMSO-d<sub>6</sub>).**



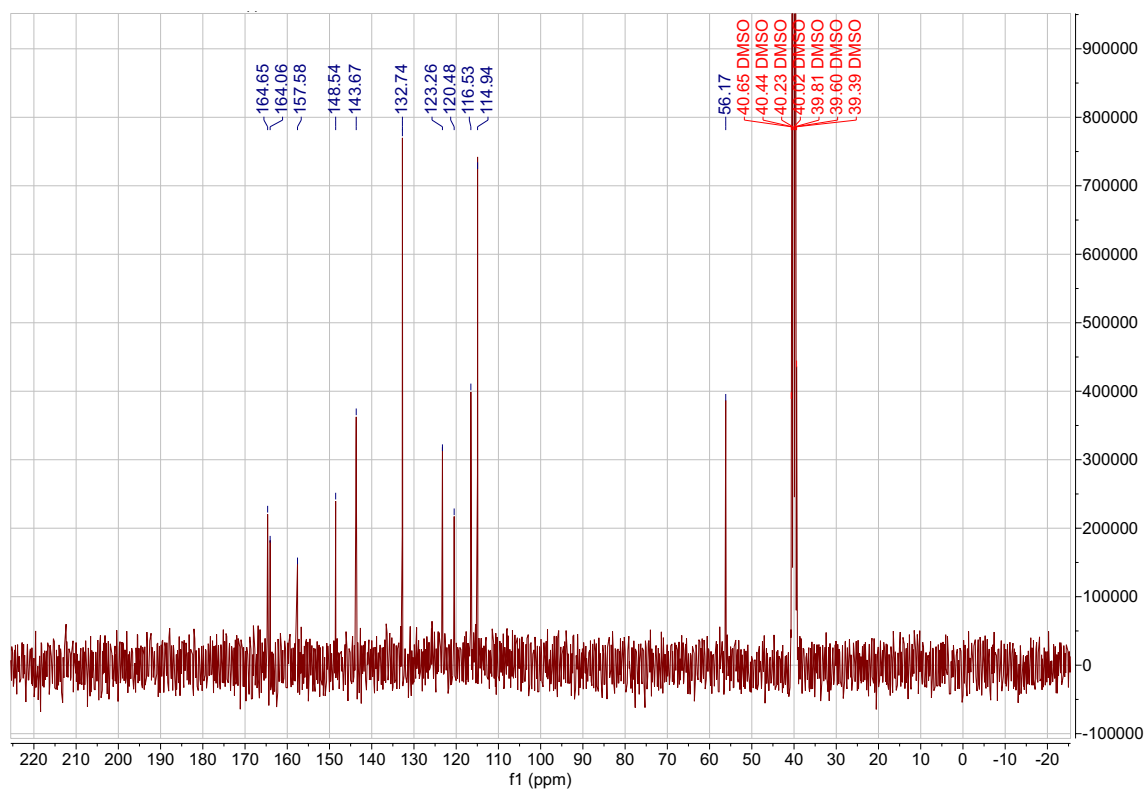
Supplementary Figure S6: <sup>1</sup>H NMR compound 45c (DMSO-d<sub>6</sub>).



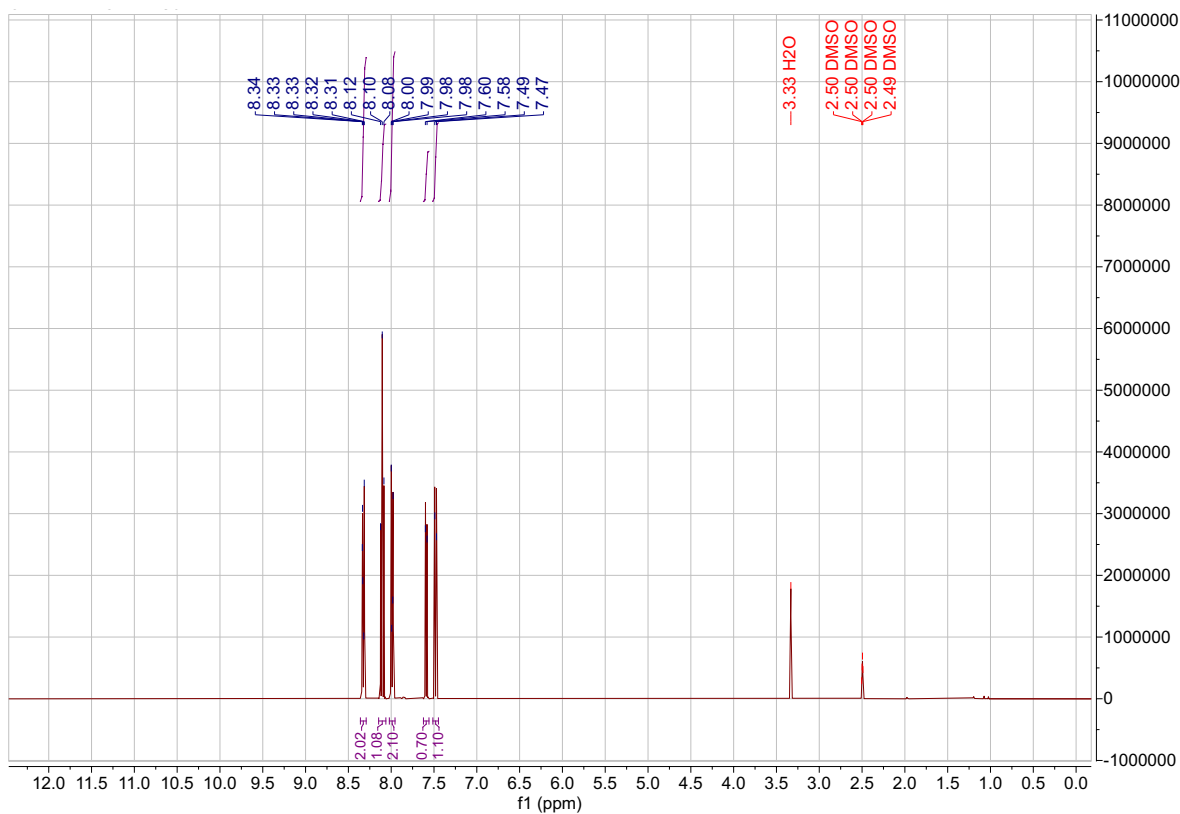
Supplementary Figure S7: <sup>13</sup>C NMR compound 45c (DMSO-d<sub>6</sub>).



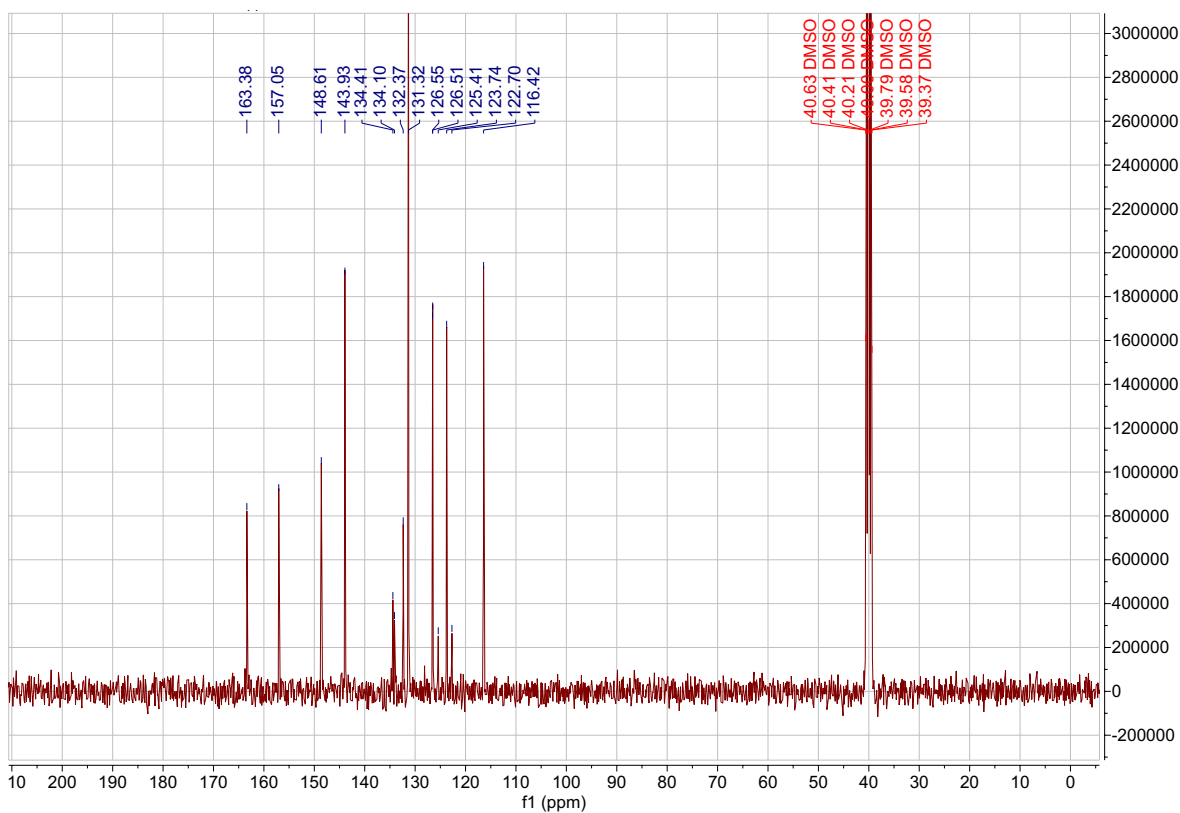
Supplementary Figure S8: <sup>1</sup>H NMR compound 45d (DMSO-d<sub>6</sub>).



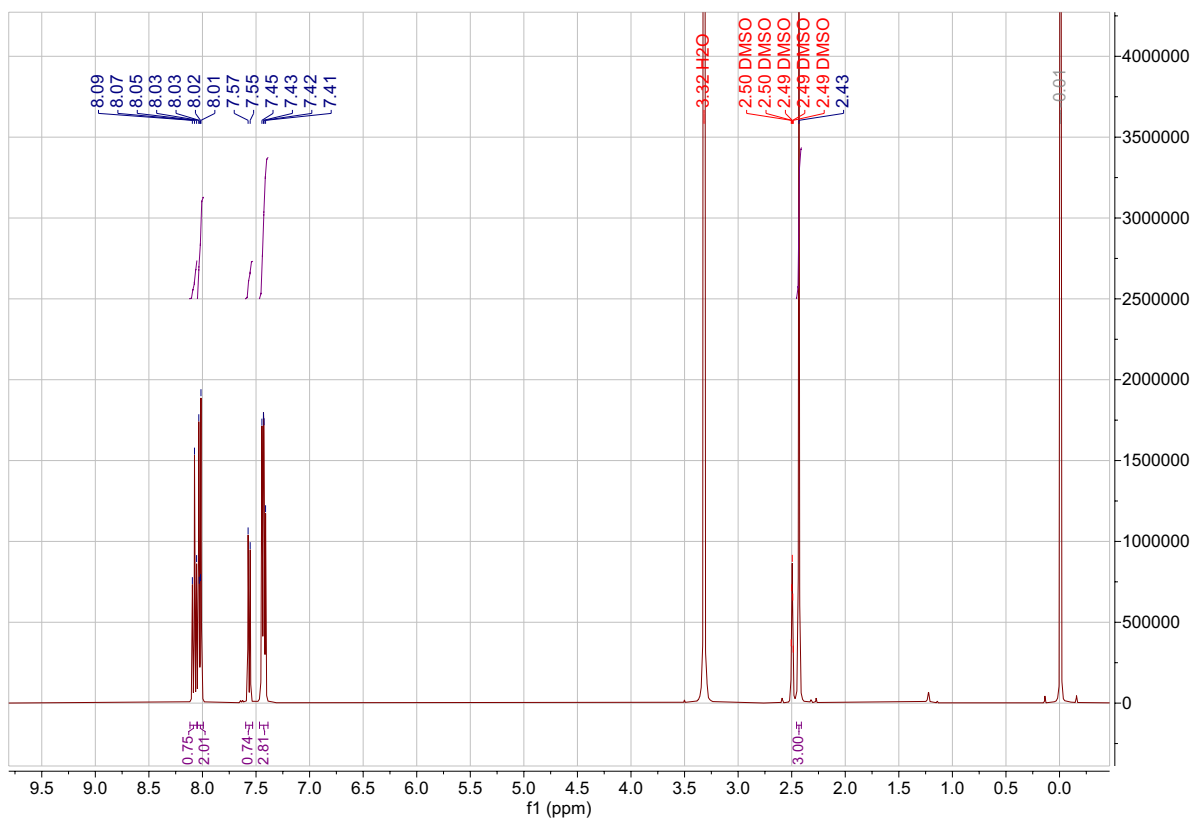
Supplementary Figure S9: <sup>13</sup>C NMR compound 45d (DMSO-d<sub>6</sub>).



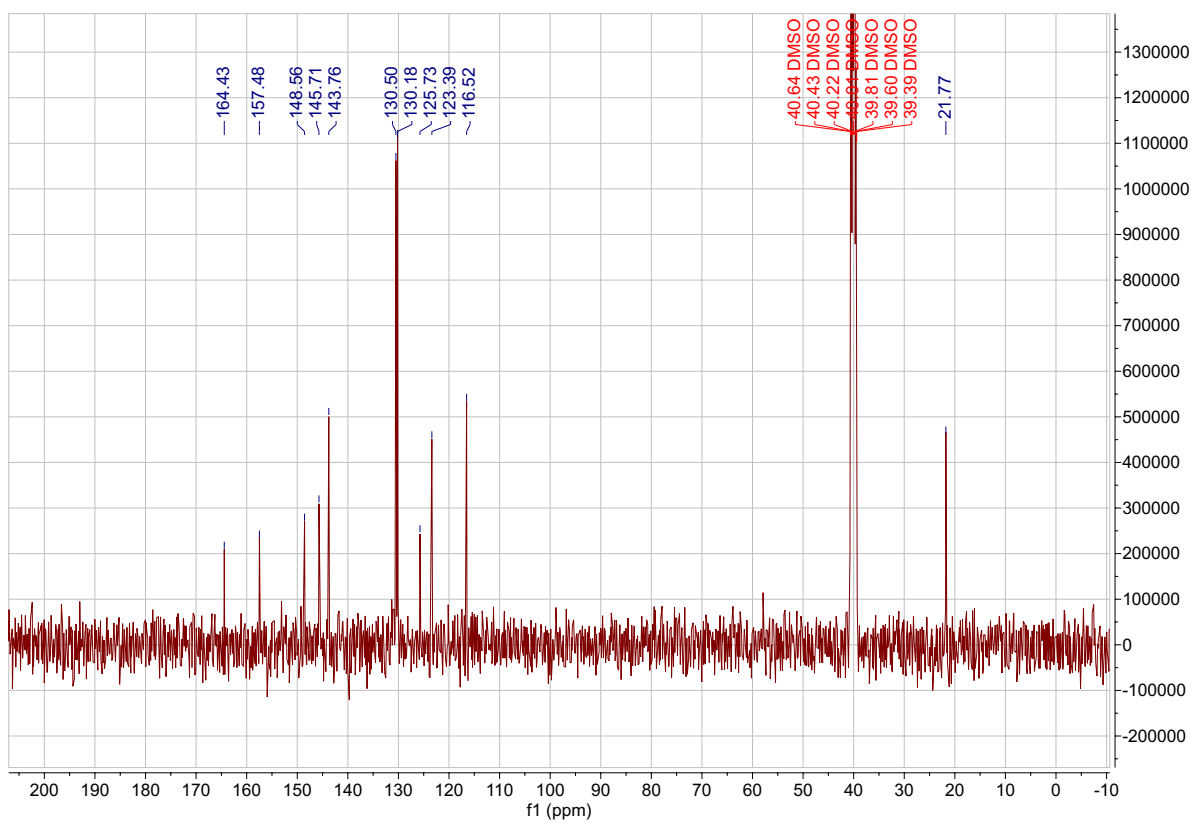
Supplementary Figure S10:  $^1\text{H}$ NMR compound 45e (DMSO- $d_6$ ).



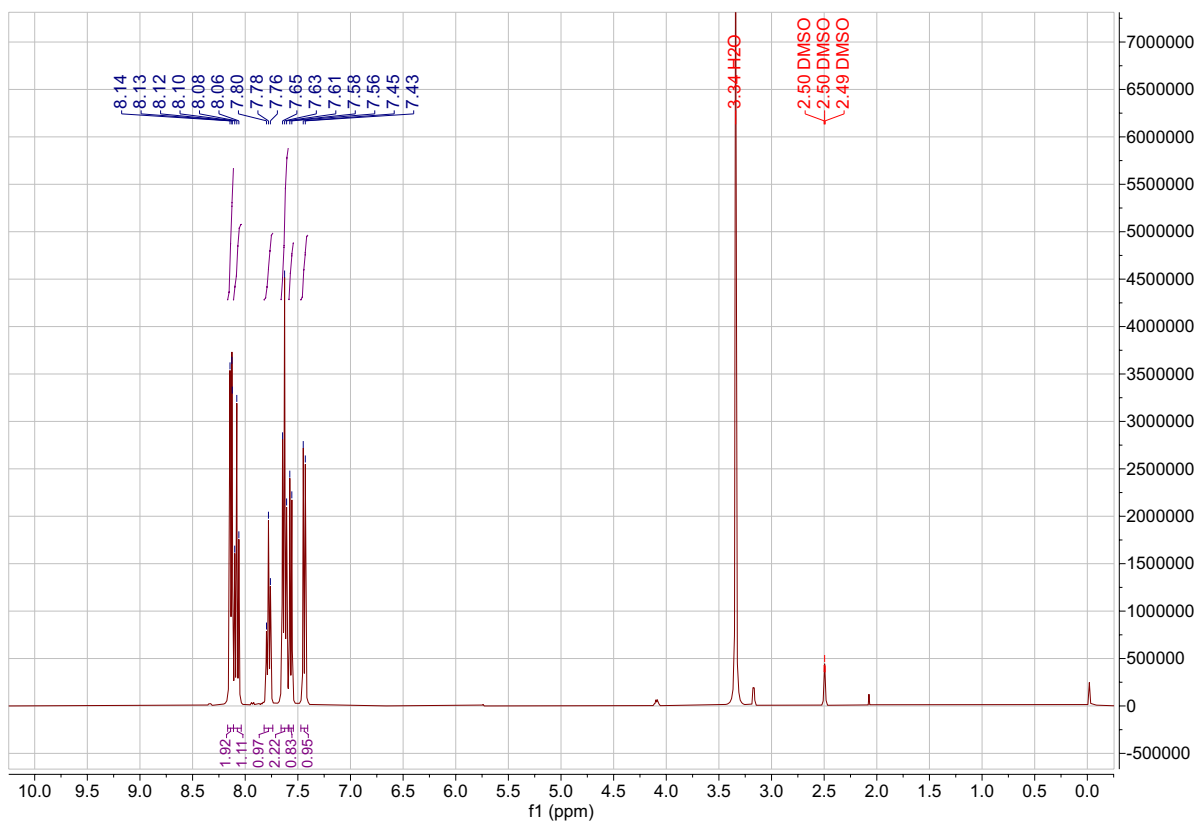
Supplementary Figure S11:  $^{13}\text{C}$ NMR compound 45e (DMSO- $d_6$ ).



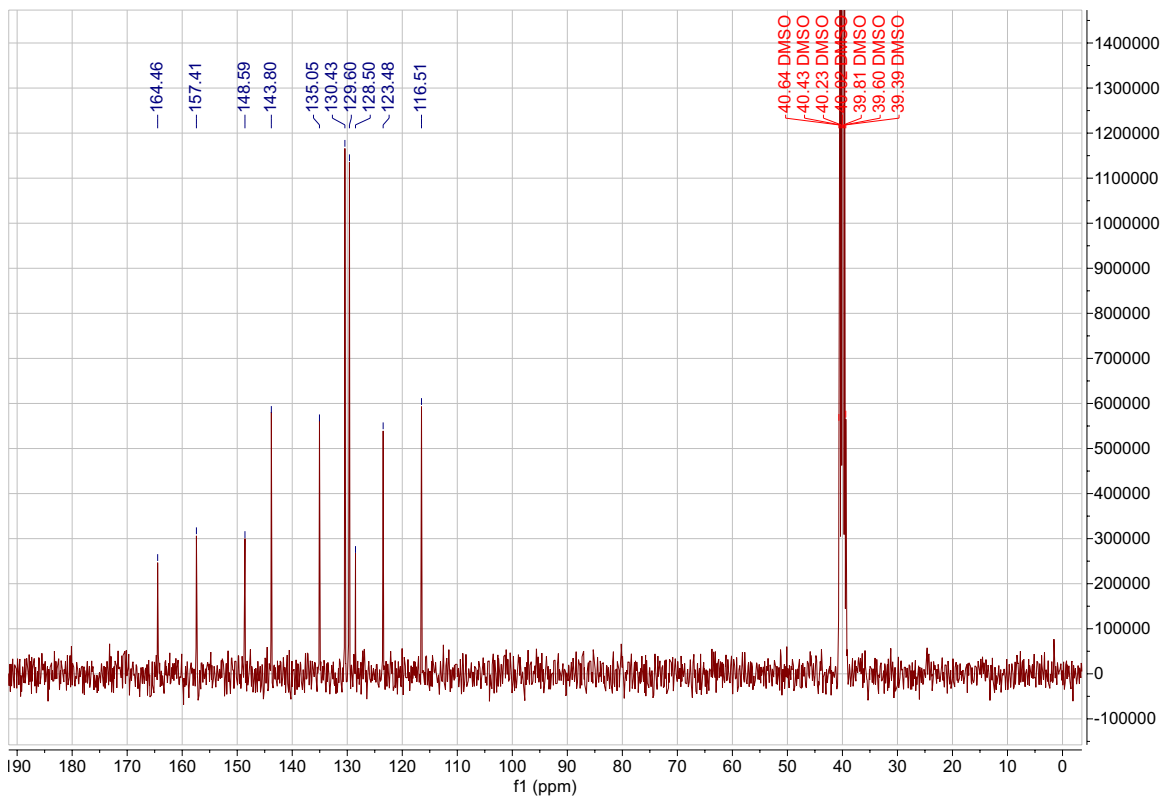
Supplementary Figure S12: <sup>1</sup>H NMR compound 45f (DMSO-d<sub>6</sub>).



Supplementary Figure S13: <sup>13</sup>C NMR compound 45f (DMSO-d<sub>6</sub>).

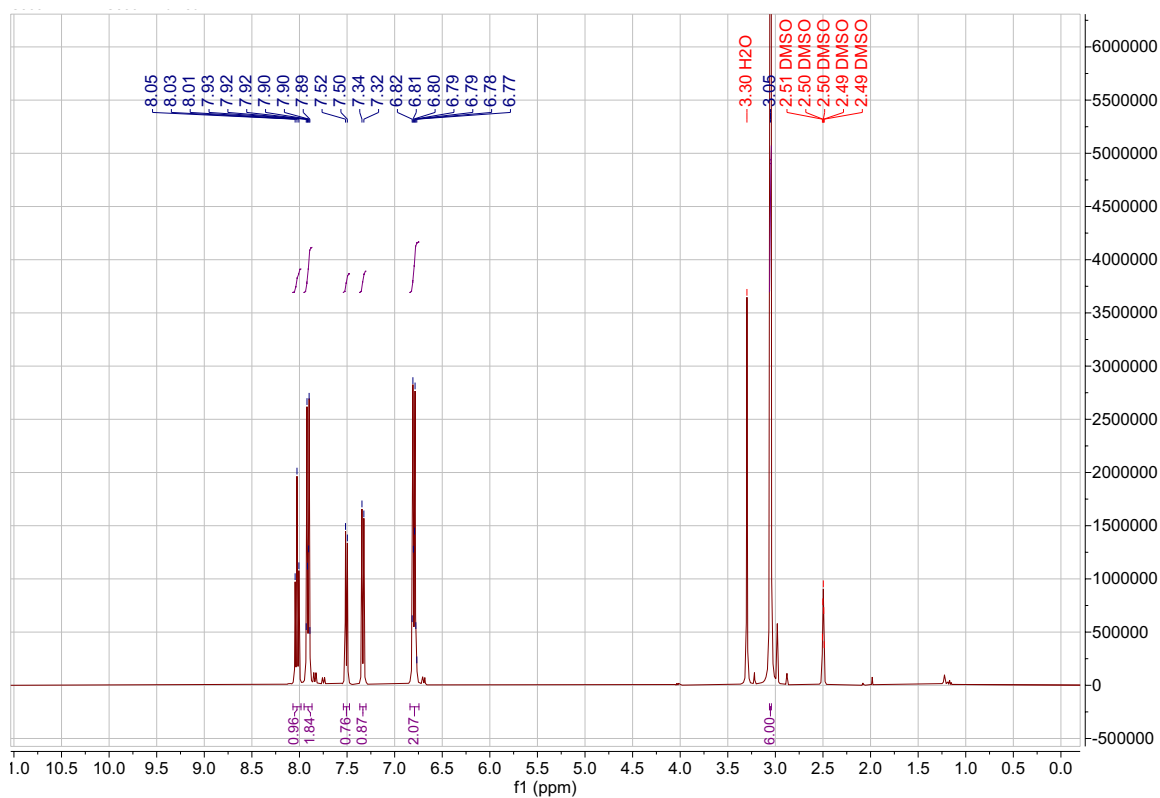


Supplementary Figure S14: <sup>1</sup>H NMR compound 45g (DMSO-d<sub>6</sub>).

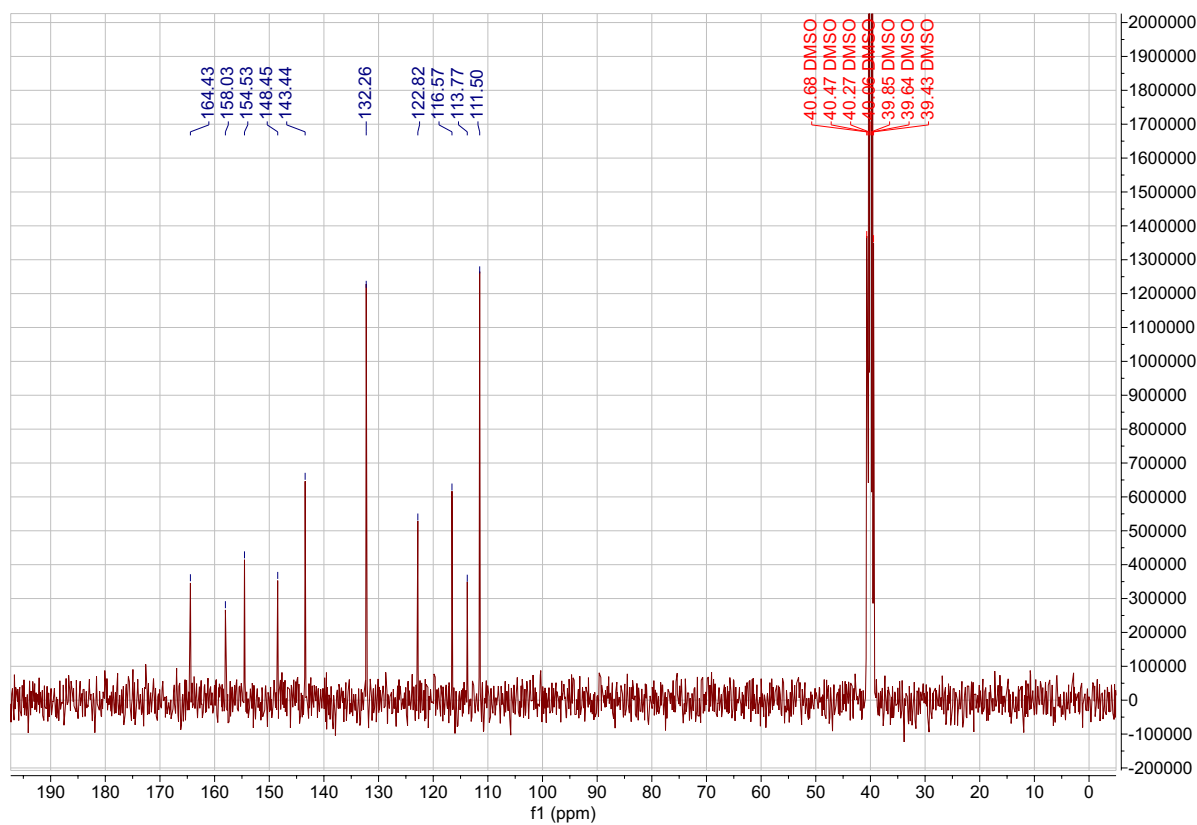


Supplementary Figure S15: <sup>13</sup>C NMR compound 45g (DMSO-d<sub>6</sub>).

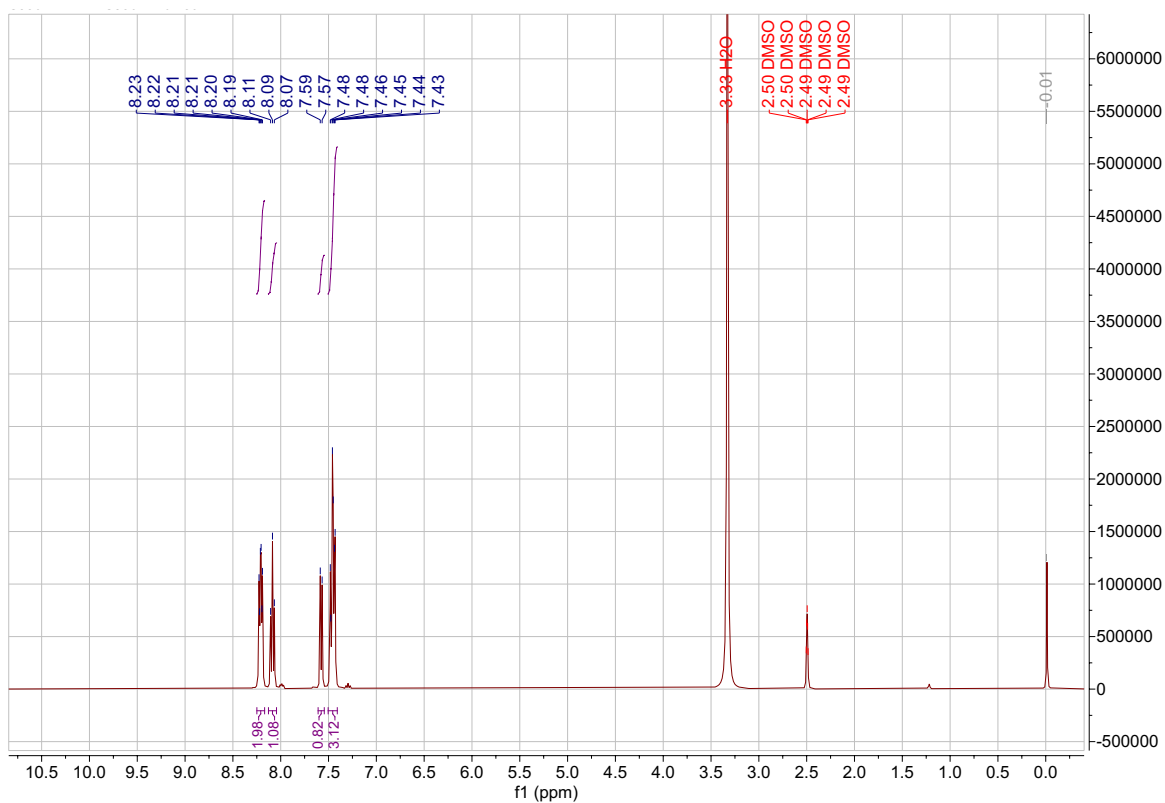




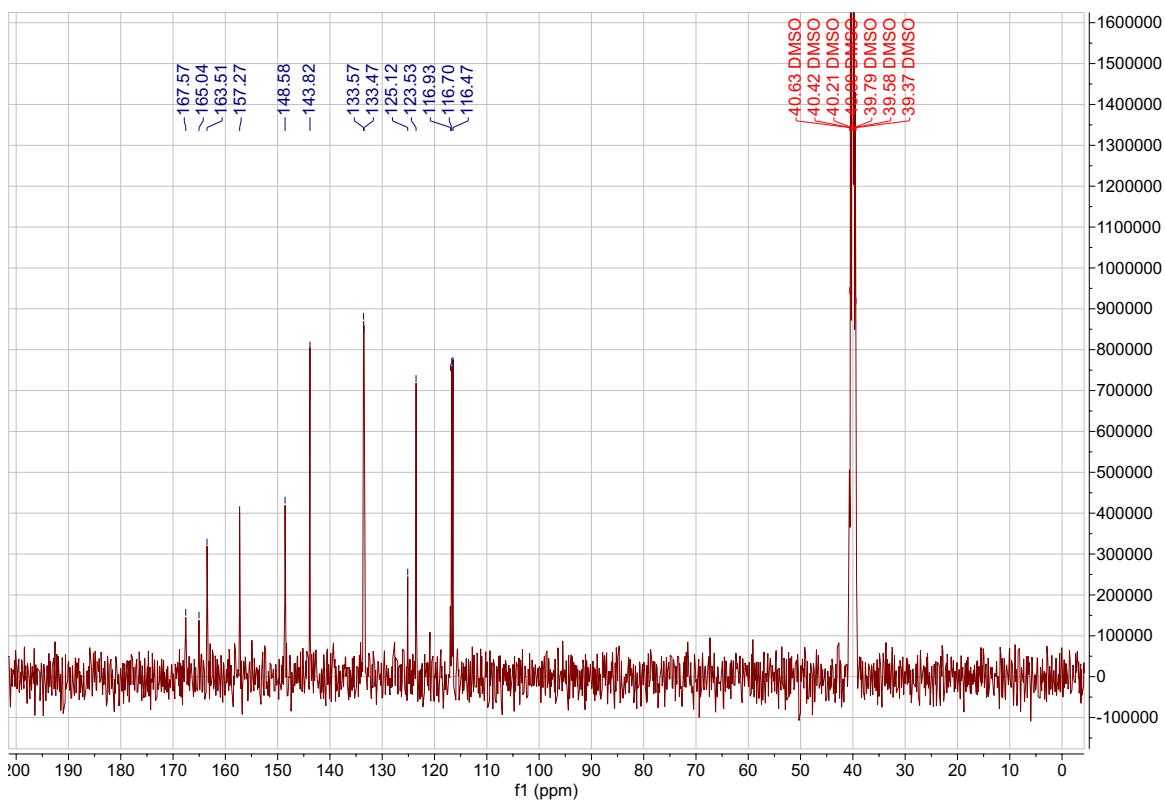
Supplementary Figure S16:  $^1\text{H}$ NMR compound 45h (DMSO- $d_6$ ).



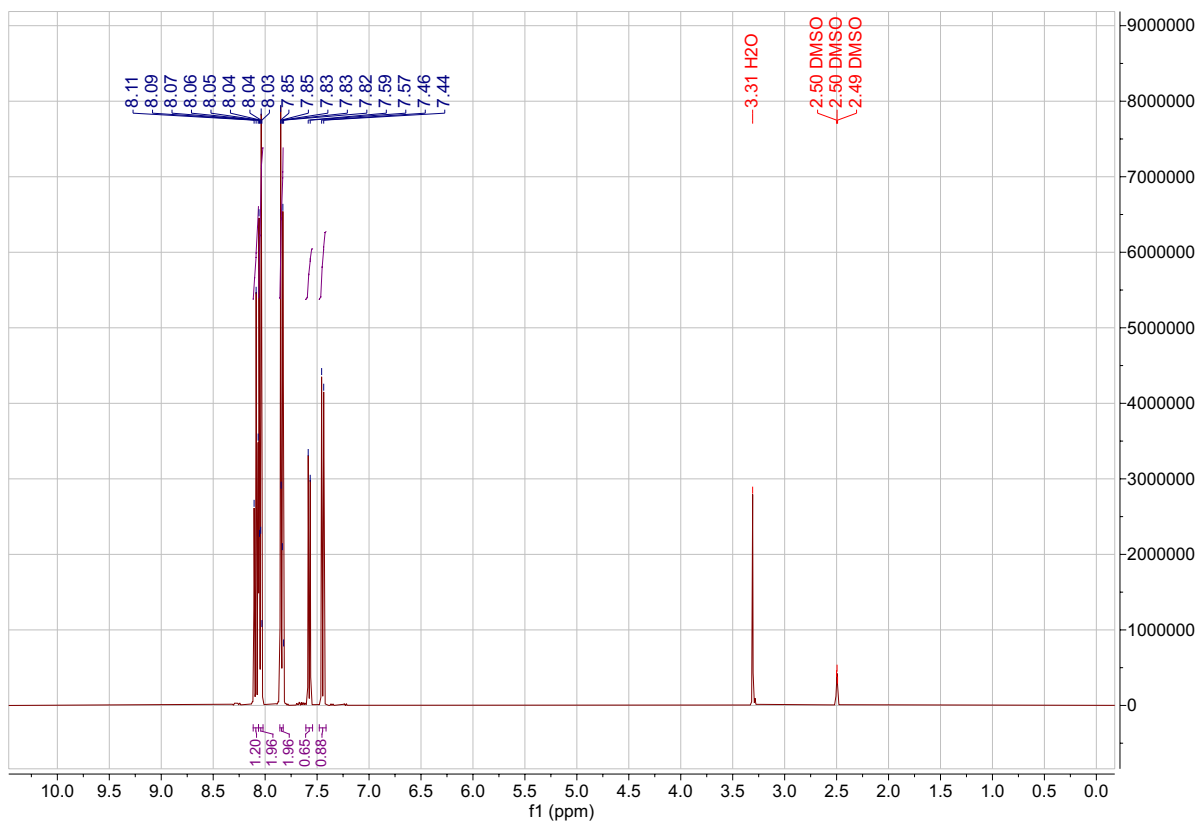
Supplementary Figure S17:  $^{13}\text{C}$ NMR compound 45h (DMSO- $d_6$ ).



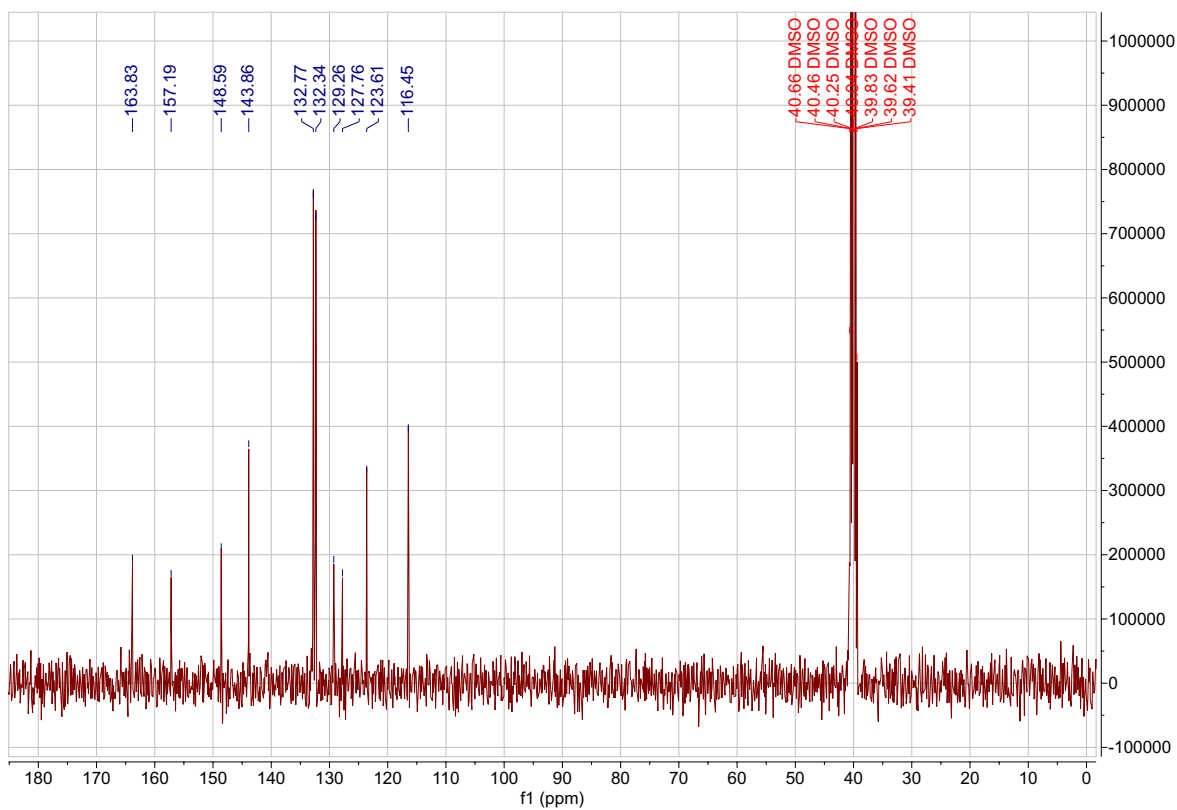
Supplementary Figure S18:  $^1\text{H}$ NMR compound 45i (DMSO- $d_6$ ).



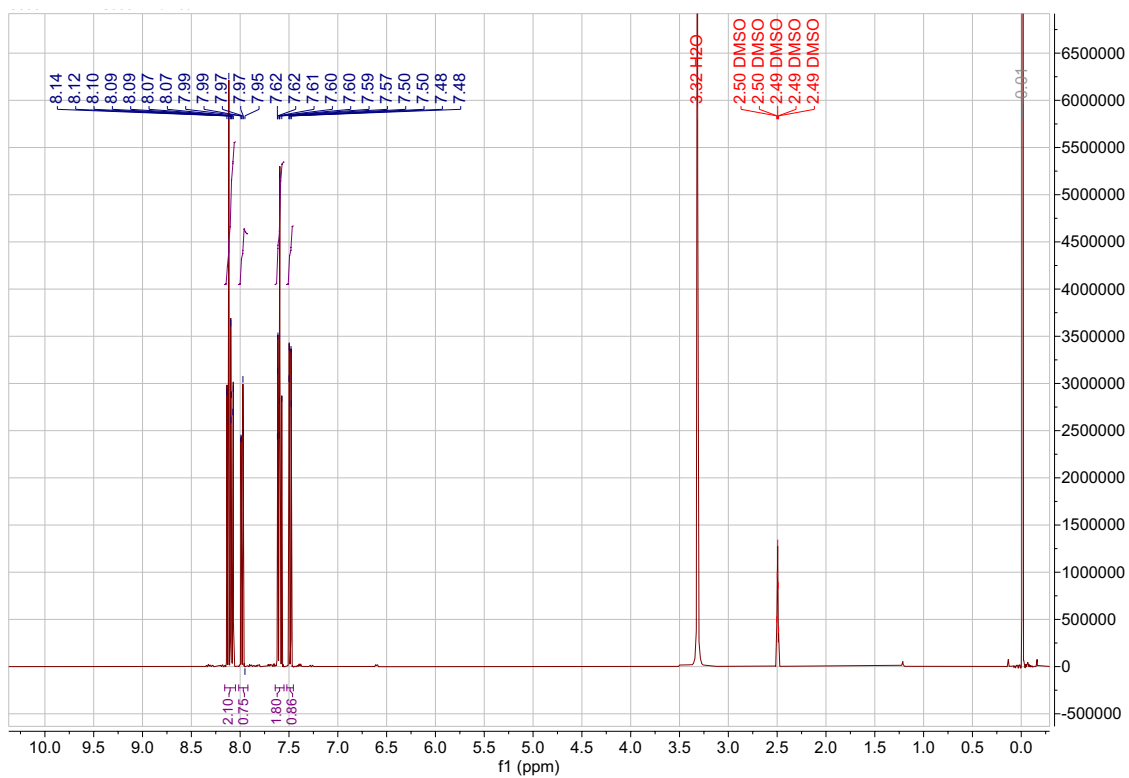
Supplementary Figure S19:  $^{13}\text{C}$ NMR compound 45i (DMSO- $d_6$ ).



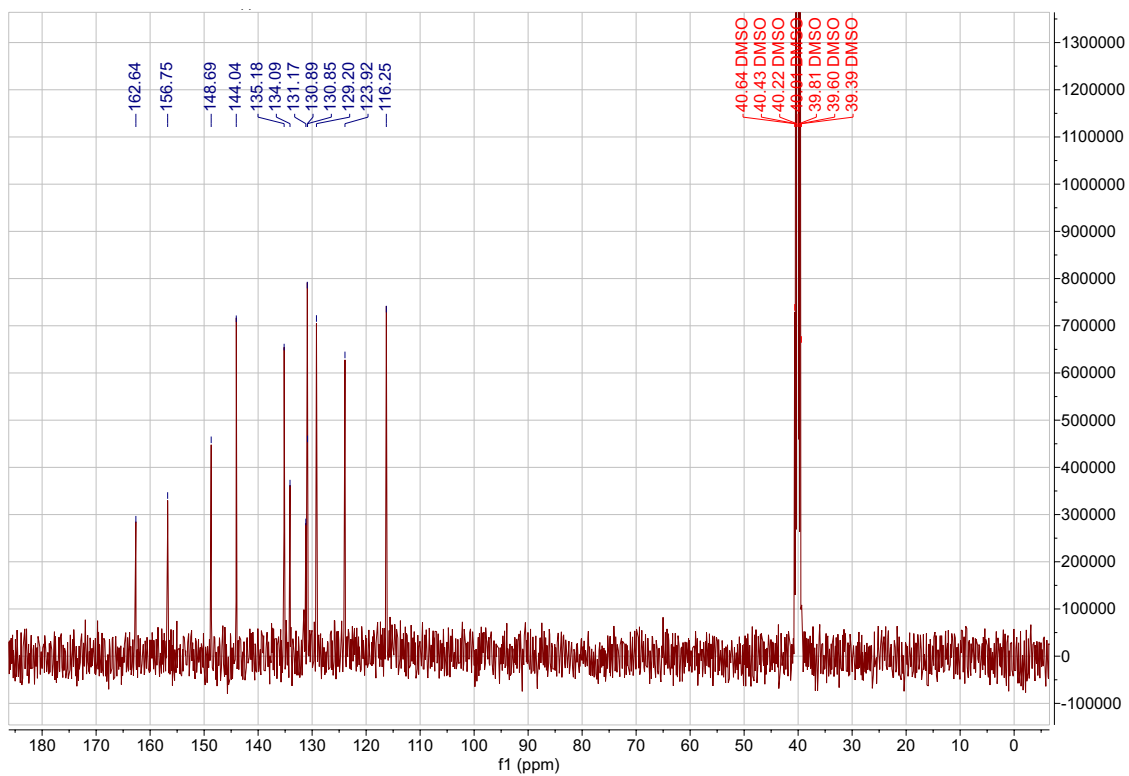
Supplementary Figure S20:  $^1\text{H}$ NMR compound 45j (DMSO- $d_6$ ).



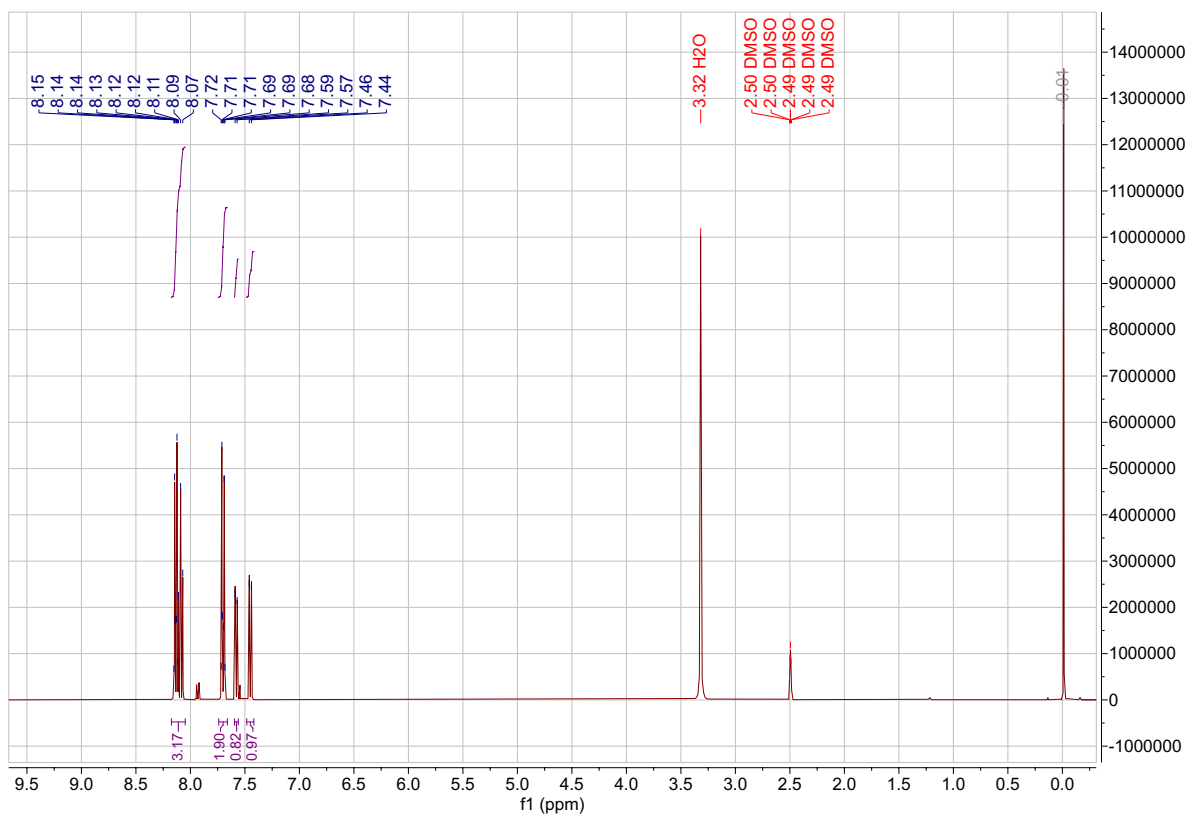
Supplementary Figure S21:  $^{13}\text{C}$ NMR compound 45j (DMSO- $d_6$ ).



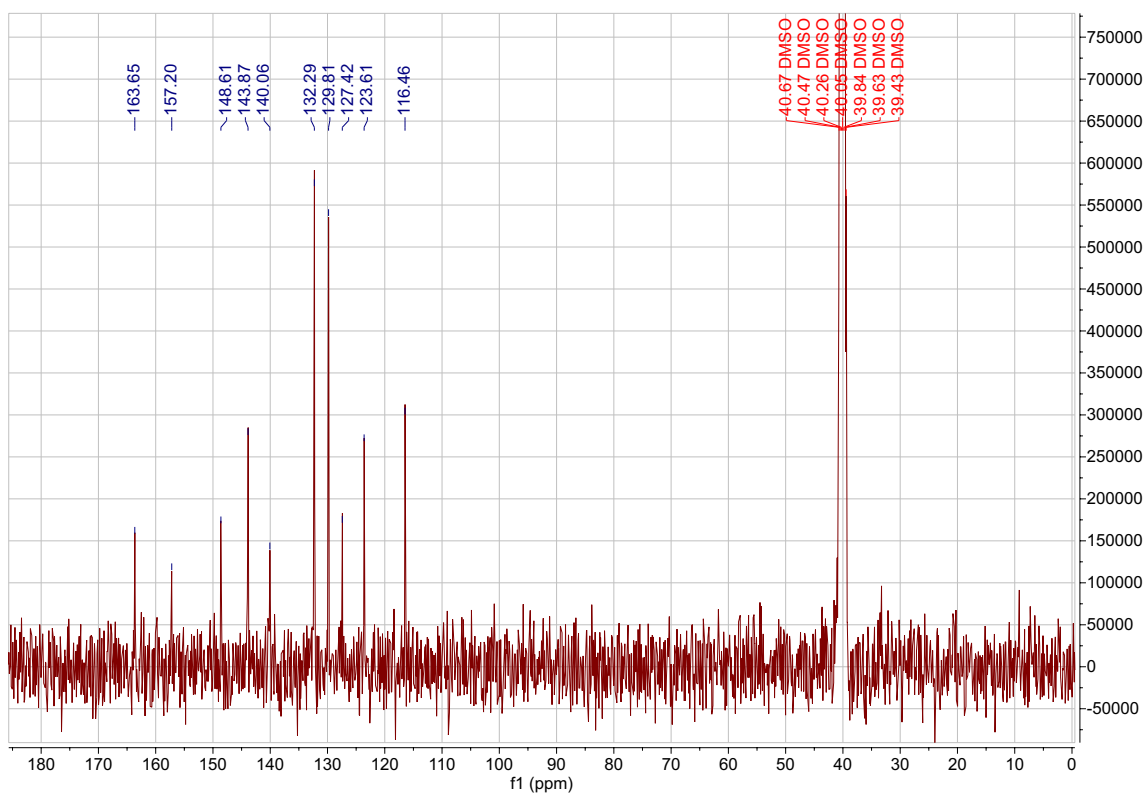
Supplementary Figure S22: <sup>1</sup>H NMR compound 45k (DMSO-d<sub>6</sub>).



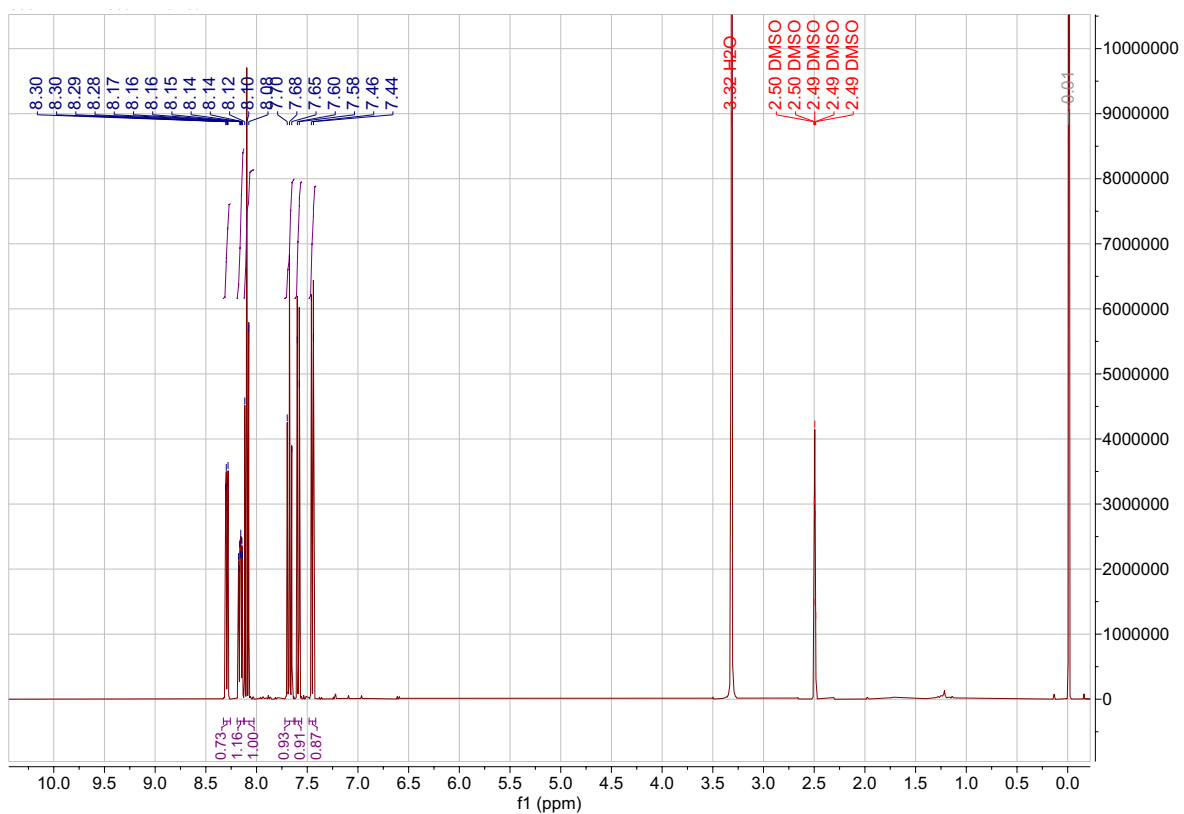
Supplementary Figure S23: <sup>13</sup>C NMR compound 45k (DMSO-d<sub>6</sub>).



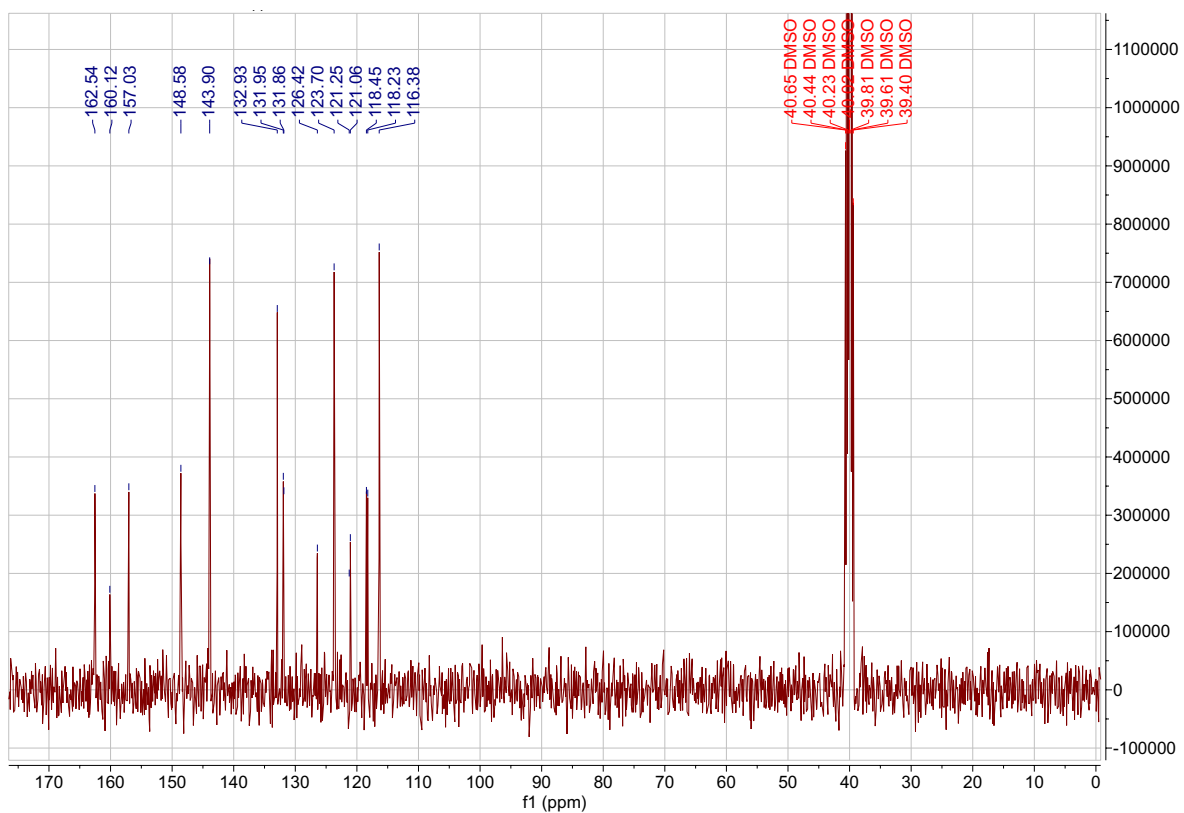
Supplementary Figure S24: <sup>1</sup>H NMR compound 45I (DMSO-d<sub>6</sub>).



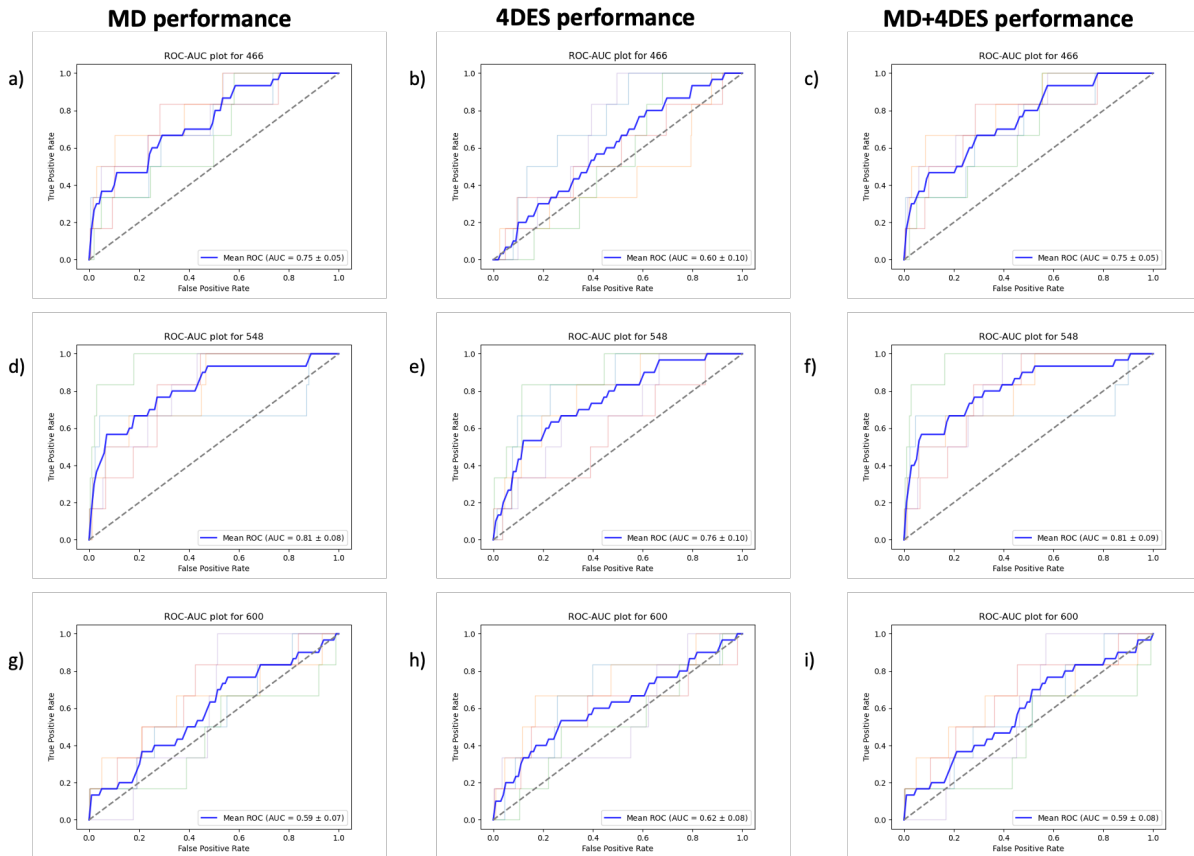
Supplementary Figure S25: <sup>13</sup>C NMR compound 45I (DMSO-d<sub>6</sub>).



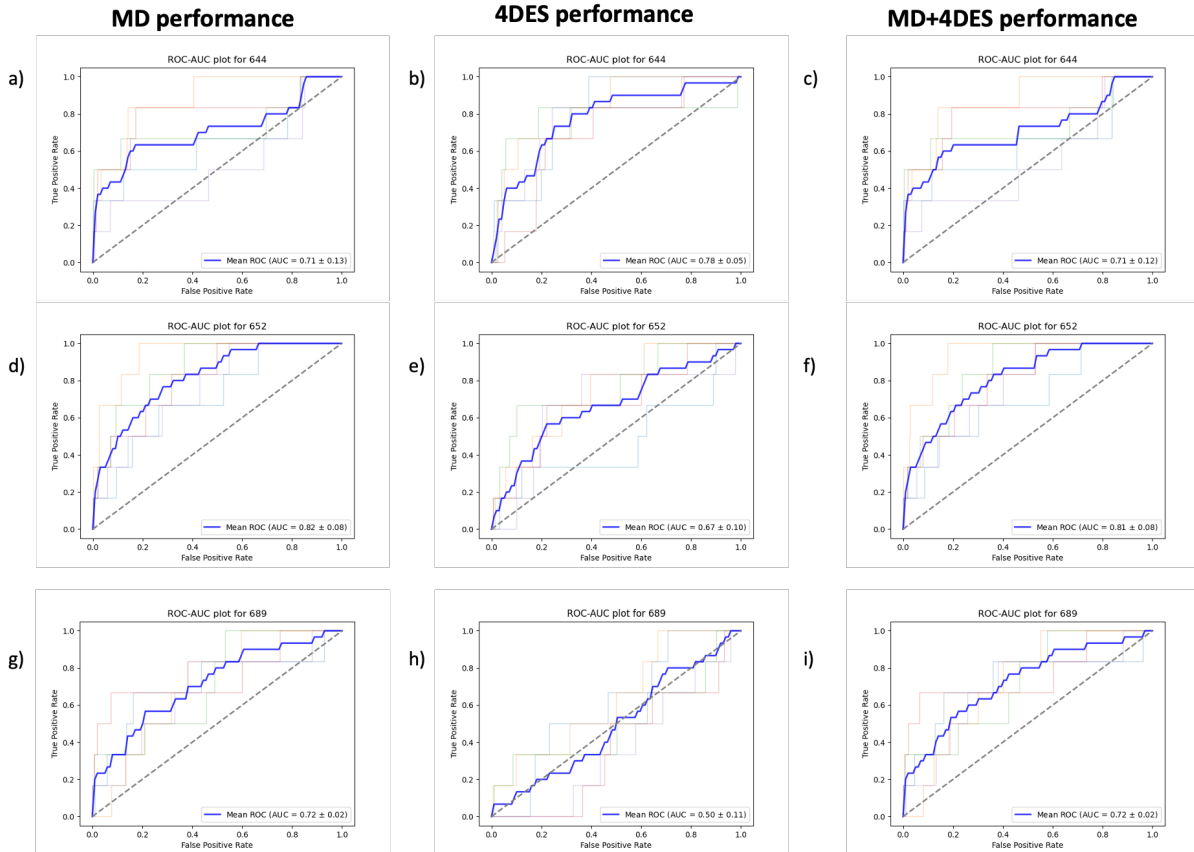
Supplementary Figure S26: <sup>1</sup>H NMR compound 45m (DMSO-d<sub>6</sub>).



Supplementary Figure S27: <sup>13</sup>C NMR compound 45m (DMSO-d<sub>6</sub>).

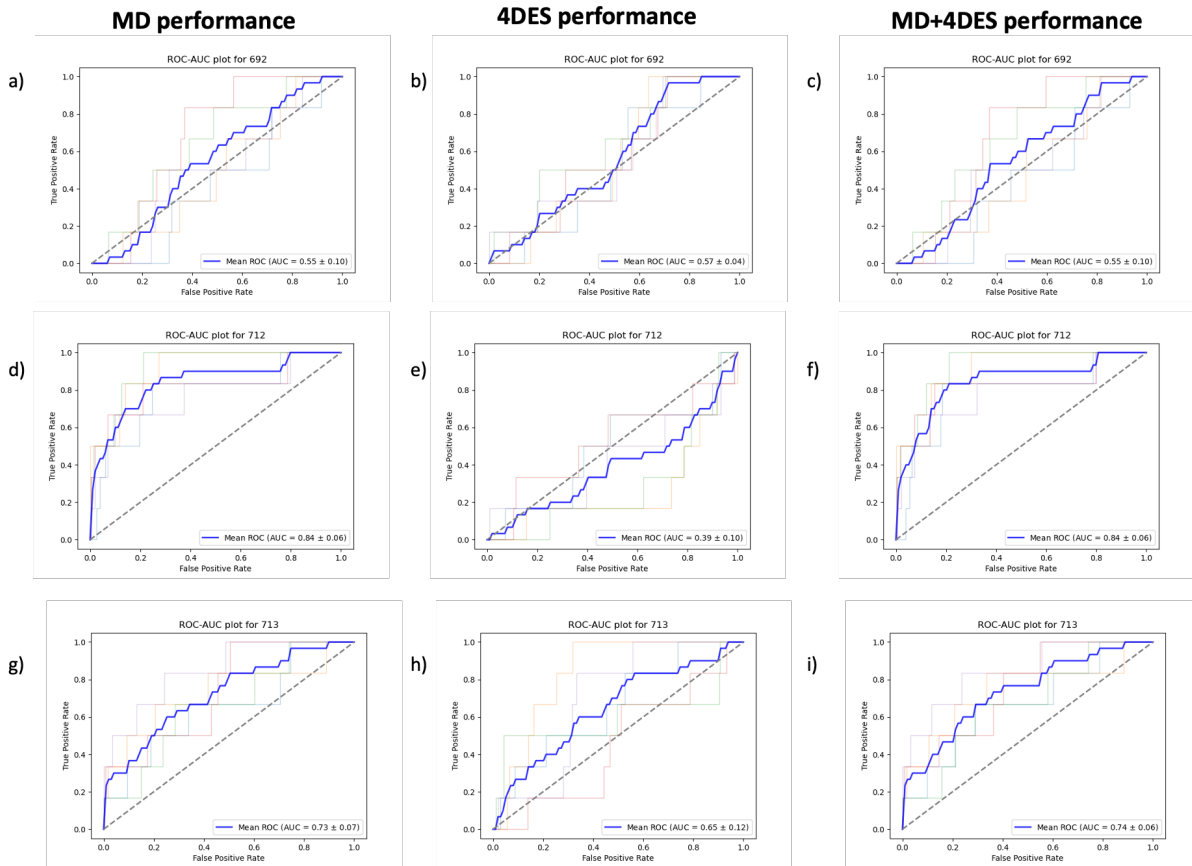


**Supplementary Figure S28.** (a) ROC-AUC for target MUV\_466 when using the MD approach; (b) ROC-AUC for target MUV\_466 when using the 4DES approach; (c) ROC-AUC for target MUV\_466 when using the MD+4DES combined approach; (d) ROC-AUC for target MUV\_548 when using the MD approach; (e) ROC-AUC for target MUV\_548 when using 4DES approach; (f) ROC-AUC for target MUV\_548 when the using the MD+4DES combined approach; (g) ROC-AUC for target MUV\_600 when using the MD approach; (h) ROC-AUC for target MUV\_600 when using the 4DES approach; (i) ROC-AUC for target MUV\_600 when using the MD+4DES combined approach.

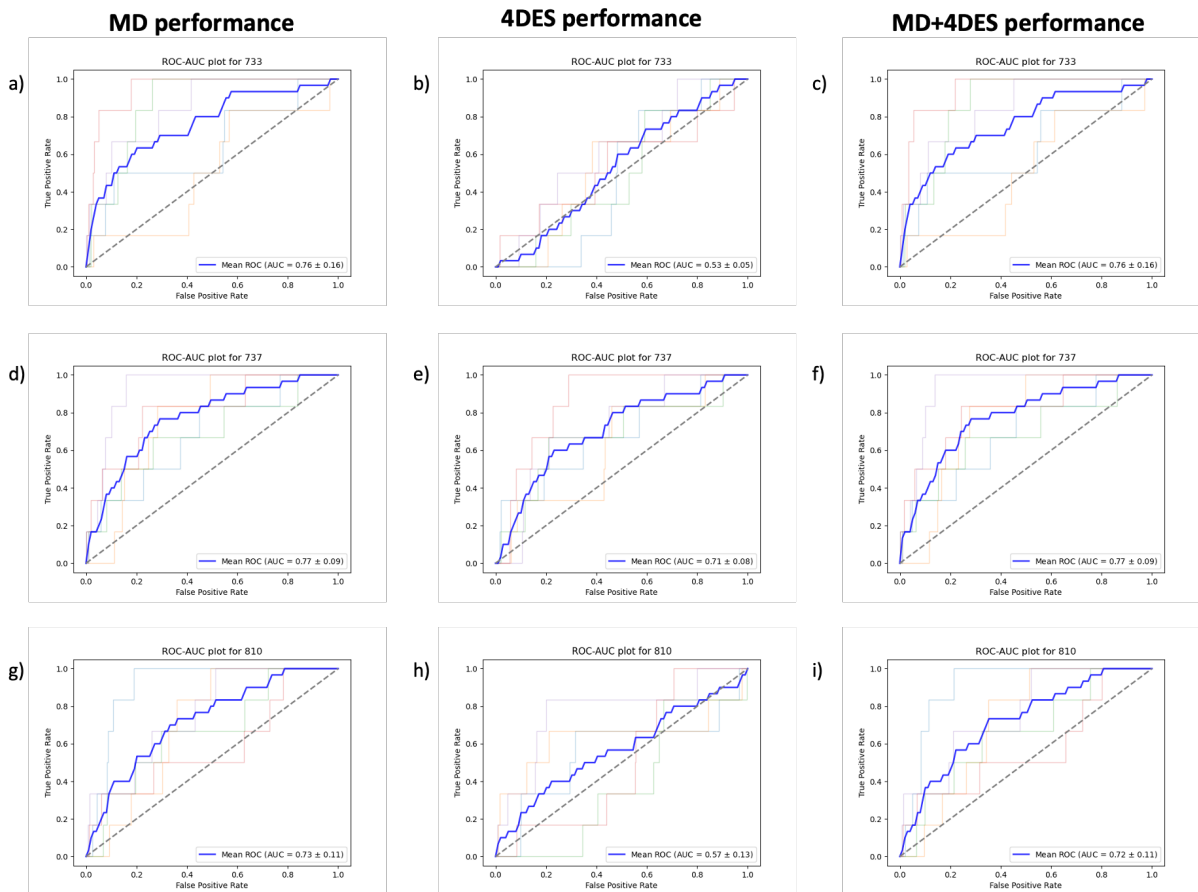


**Supplementary Figure S29.** (a) ROC-AUC for target MUV\_644 when using the MD approach; (b) ROC-AUC for target MUV\_644 when using the 4DES approach; (c) ROC-AUC for target MUV\_644 when using the MD+4DES combined approach; (d) ROC-AUC for target MUV\_652 when using the MD approach; (e) ROC-AUC for target MUV\_652 when using 4DES approach; (f) ROC-AUC for target MUV\_652 when the using the MD+4DES combined approach; (g) ROC-AUC for target MUV\_689 when using the MD approach; (h) ROC-AUC for target MUV\_689 when using the 4DES approach; (i) ROC-AUC for target MUV\_689 when using the MD+4DES combined approach.

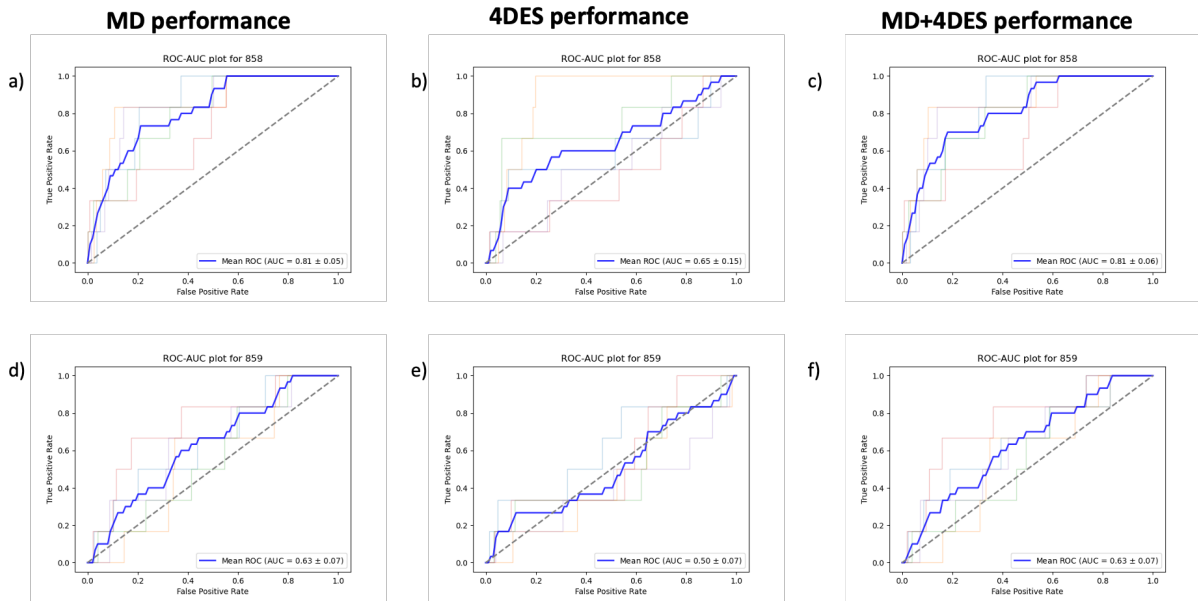




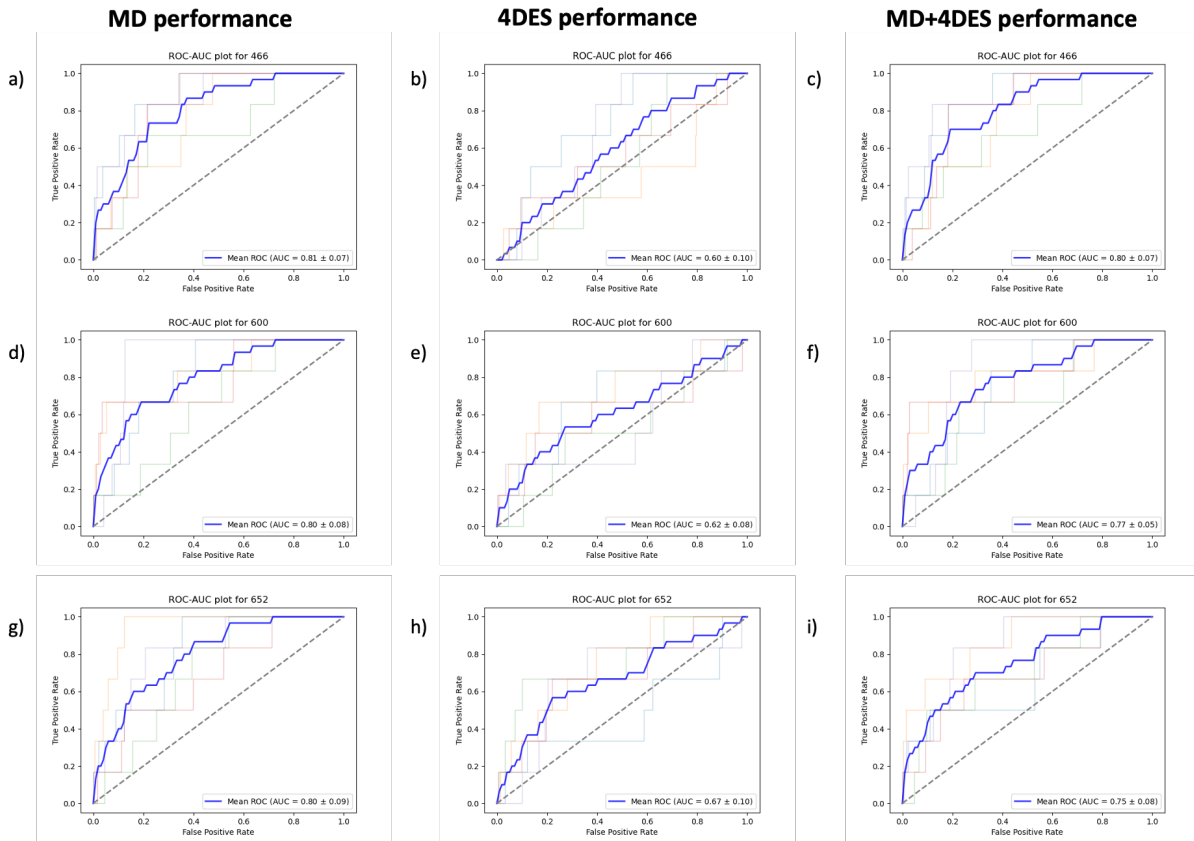
**Supplementary Figure S30.** (a) ROC-AUC for target MUV\_692 when using the MD approach; (b) ROC-AUC for target MUV\_692 when using the 4DES approach; (c) ROC-AUC for target MUV\_692 when using the MD+4DES combined approach; (d) ROC-AUC for target MUV\_712 when using the MD approach; (e) ROC-AUC for target MUV\_712 when using 4DES approach; (f) ROC-AUC for target MUV\_712 when the using the MD+4DES combined approach; (g) ROC-AUC for target MUV\_713 when using the MD approach; (h) ROC-AUC for target MUV\_713 when using the 4DES approach; (i) ROC-AUC for target MUV\_713 when using the MD+4DES combined approach.



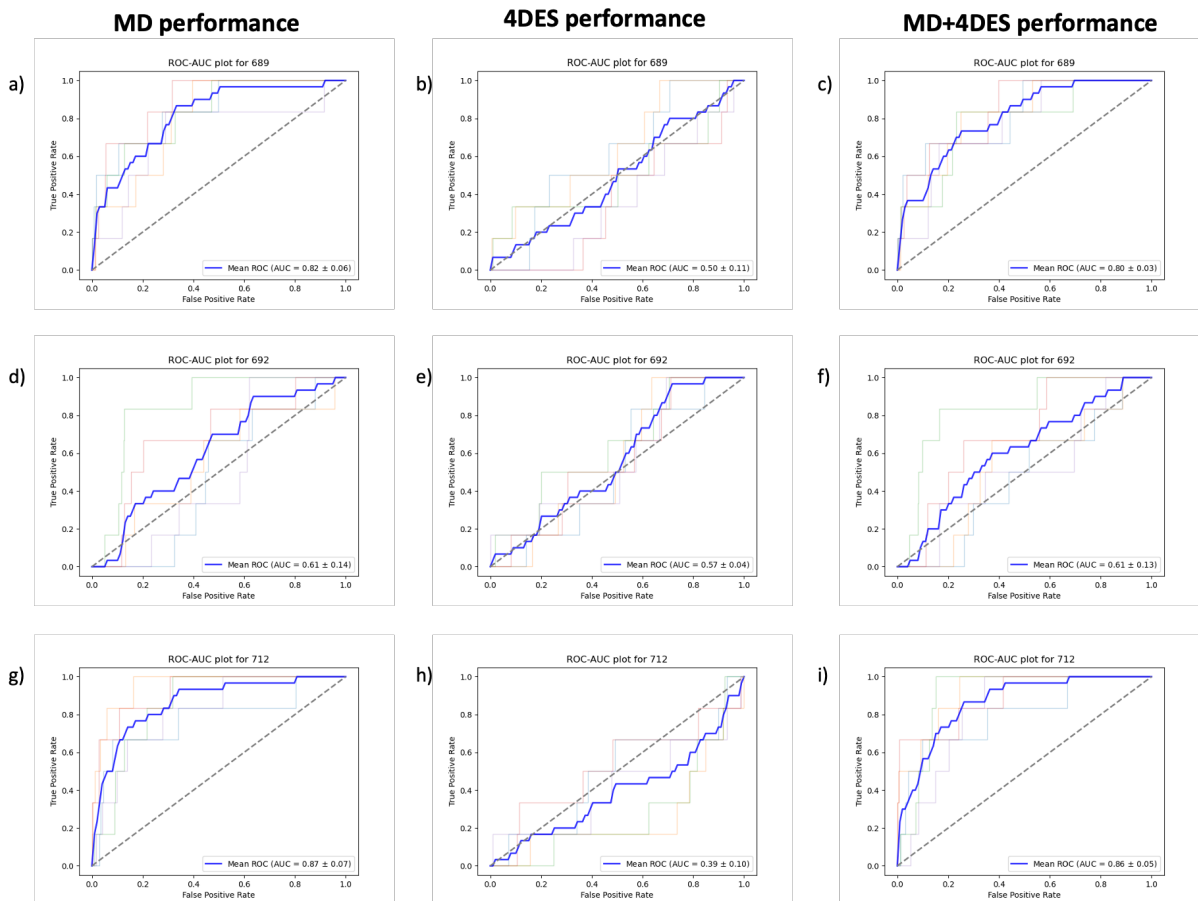
**Supplementary Figure S31.** (a) ROC-AUC for target MUV\_733 when using the MD approach; (b) ROC-AUC for target MUV\_733 when using the 4DES approach; (c) ROC-AUC for target MUV\_733 when using the MD+4DES combined approach; (d) ROC-AUC for target MUV\_737 when using the MD approach; (e) ROC-AUC for target MUV\_737 when using 4DES approach; (f) ROC-AUC for target MUV\_737 when the using the MD+4DES combined approach; (g) ROC-AUC for target MUV\_810 when using the MD approach; (h) ROC-AUC for target MUV\_810 when using the 4DES approach; (i) ROC-AUC for target MUV\_810 when using the MD+4DES combined approach.



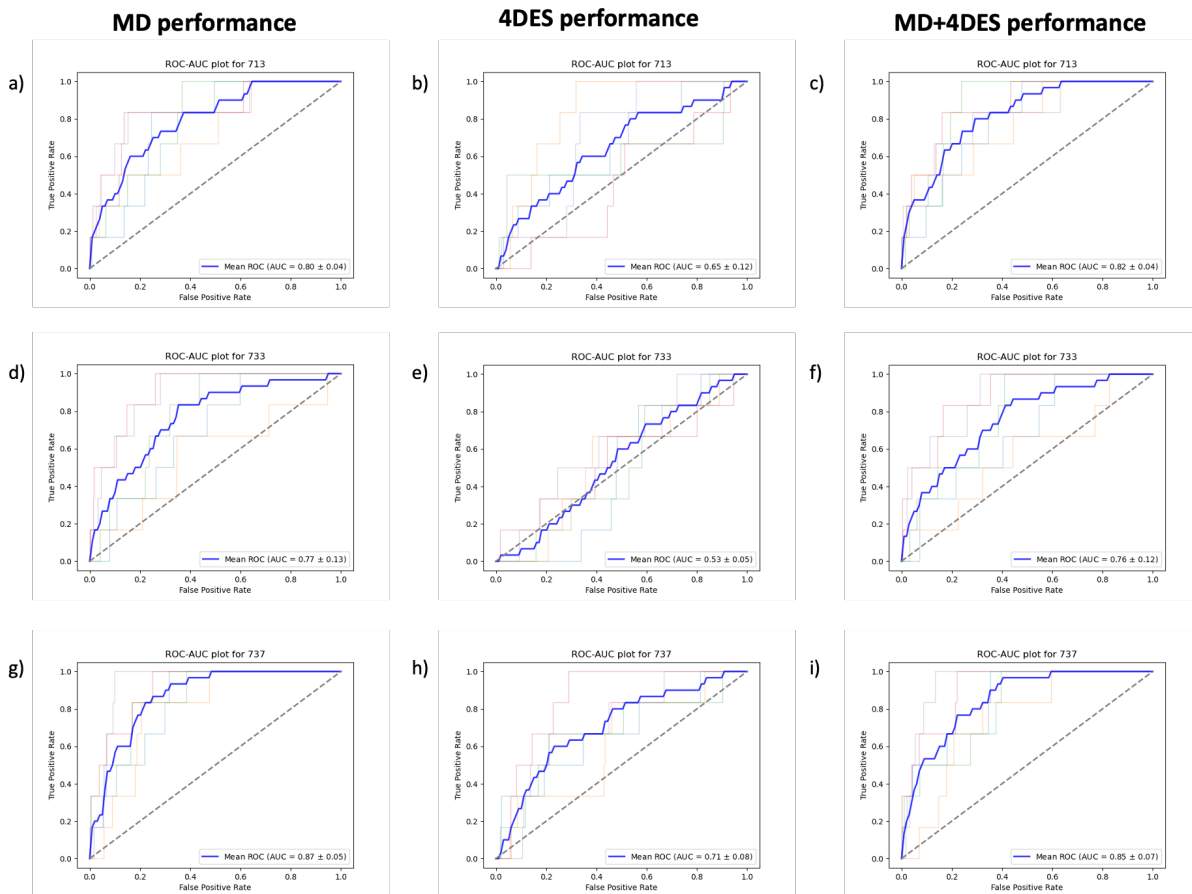
**Supplementary Figure S32.** (a) ROC-AUC for target MUV\_858 when using the MD approach; (b) ROC-AUC for target MUV\_858 when using the 4DES approach; (c) ROC-AUC for target MUV\_858 when using the MD+4DES combined approach; (d) ROC-AUC for target MUV\_859 when using the MD approach; (e) ROC-AUC for target MUV\_859 when using 4DES approach; (f) ROC-AUC for target MUV\_859 when the using the MD+4DES combined approach.



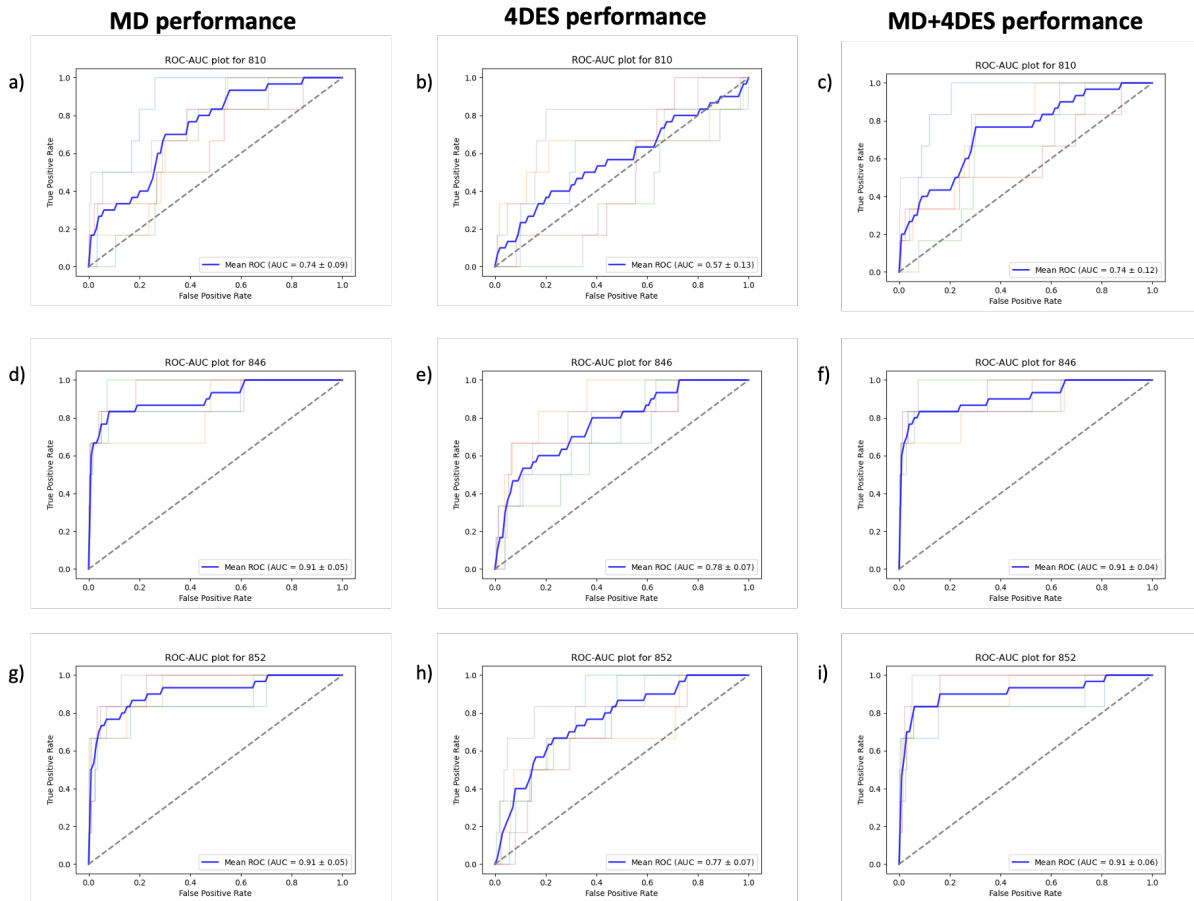
**Supplementary Figure S33. (a)** ROC-AUC for target MUV\_466 when using the MD approach; **(b)** ROC-AUC for target MUV\_466 when using the 4DES approach; **(c)** ROC-AUC for target MUV\_466 when using the MD+4DES combined approach; **(d)** ROC-AUC for target MUV\_600 when using the MD approach; **(e)** ROC-AUC for target MUV\_600 when using 4DES approach; **(f)** ROC-AUC for target MUV\_600 when the using the MD+4DES combined approach; **(g)** ROC-AUC for target MUV\_652 when using the MD approach; **(h)** ROC-AUC for target MUV\_652 when using the 4DES approach; **(i)** ROC-AUC for target MUV\_652 when using the MD+4DES combined approach.



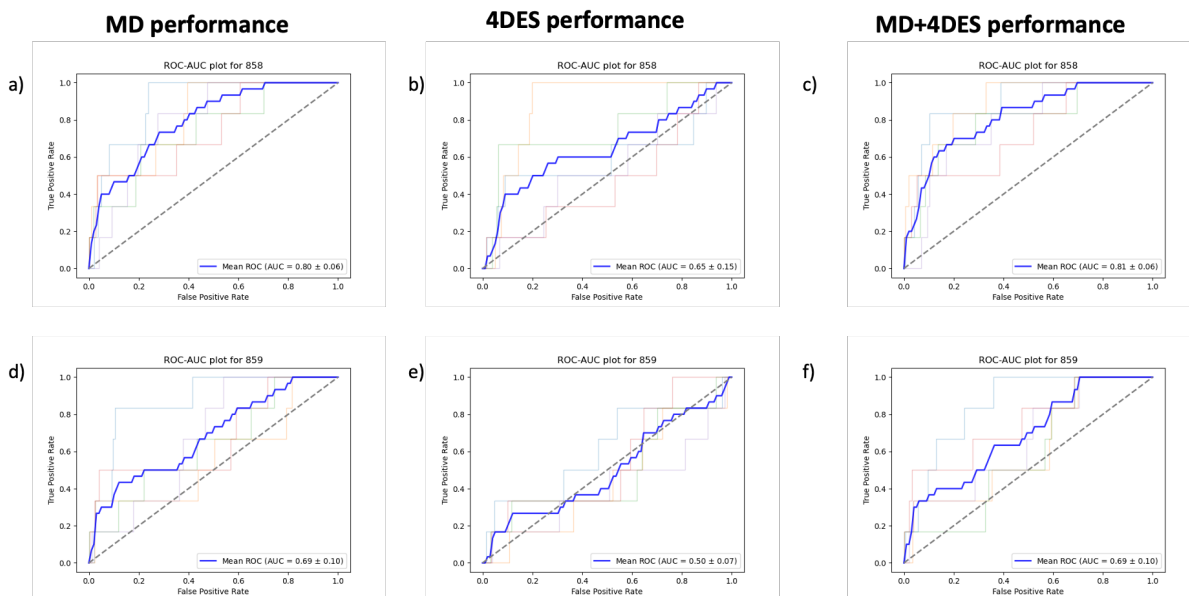
**Supplementary Figure S34. (a)** ROC-AUC for target MUV\_689 when using the MD approach; **(b)** ROC-AUC for target MUV\_689 when using the 4DES approach; **(c)** ROC-AUC for target MUV\_689 when using the MD+4DES combined approach; **(d)** ROC-AUC for target MUV\_692 when using the MD approach; **(e)** ROC-AUC for target MUV\_692 when using 4DES approach; **(f)** ROC-AUC for target MUV\_692 when the using the MD+4DES combined approach; **(g)** ROC-AUC for target MUV\_712 when using the MD approach; **(h)** ROC-AUC for target MUV\_712 when using the 4DES approach; **(i)** ROC-AUC for target MUV\_712 when using the MD+4DES combined approach.



**Supplementary Figure S35.** (a) ROC-AUC for target MUV\_713 when using the MD approach; (b) ROC-AUC for target MUV\_713 when using the 4DES approach; (c) ROC-AUC for target MUV\_713 when using the MD+4DES combined approach; (d) ROC-AUC for target MUV\_733 when using the MD approach; (e) ROC-AUC for target MUV\_733 when using 4DES approach; (f) ROC-AUC for target MUV\_733 when the using the MD+4DES combined approach; (g) ROC-AUC for target MUV\_737 when using the MD approach; (h) ROC-AUC for target MUV\_737 when using the 4DES approach; (i) ROC-AUC for target MUV\_737 when using the MD+4DES combined approach.



**Supplementary Figure S36.** (a) ROC-AUC for target MUV\_810 when using the MD approach; (b) ROC-AUC ve for target MUV\_810 when using the 4DES approach; (c) ROC-AUC for target MUV\_810 when using the MD+4DES combined approach; (d) ROC-AUC for target MUV\_846 when using the MD approach; (e) ROC-AUC for target MUV\_846 when using 4DES approach; (f) ROC-AUC for target MUV\_846 when the using the MD+4DES combined approach; (g) ROC-AUC for target MUV\_852 when using the MD approach; (h) ROC-AUC for target MUV\_852 when using the 4DES approach; (i) ROC-AUC for target MUV\_852 when using the MD+4DES combined approach.



**Supplementary Figure S37. (a)** ROC-AUC for target MUV\_858 when using the MD approach; **(b)** ROC-AUC for target MUV\_858 when using the 4DES approach; **(c)** ROC-AUC for target MUV\_858 when using the MD+4DES combined approach; **(d)** ROC-AUC for target MUV\_859 when using the MD approach; **(e)** ROC-AUC for target MUV\_859 when using 4DES approach; **(f)** ROC-AUC for target MUV\_859 when the using the MD+4DES combined approach.



## Acknowledgements

I would like to express my sincere gratitude to Prof. Antonino Lauria and Prof. Annamaria Martorana for their precious supervision, teaching, and mentorship. Their guidance has been fundamental to my academic and professional growth.

My heartfelt thanks to Dr. Gabriele La Monica for being a constant source of inspiration and a guiding light throughout this PhD journey and in everyday life.

I am grateful to my colleague, Federica Alamia for her remarkable sensitivity, kindness, and continuous support.

Thanks are due to Prof. Giuseppe Pizzolanti for *in vitro* enzymatic inhibition assays.

A special thanks to Prof. Paul W. Finn for his infinite hospitality at the Oxford Drug Design (ODD), his constant support, kindness, and niceness and for being always prompt and available to provide me suggestions.

Thanks to the entire ODD Team, with special mentions to Prof. Richard I. Cooper and Dr. Aras Asaad, for their ongoing support and their crucial help in developing the Machine Learning Script for my project.

A heartfelt thanks to Marco, Computational Chemist at ODD, who has been both an invaluable guide in my research project and a true friend, thanks for making me feel at home in Oxford. Part of the text and figures were reproduced or adapted from published manuscripts (Bono, A. et al., Targeting SARS-CoV-2 Main Protease for Treatment of COVID-19: Covalent Inhibitors Structure-Activity Relationship Insights and Evolution Perspectives, *Journal of Medicinal Chemistry* 2022, doi: 10.1021/acs.jmedchem.2c01005 for the introduction and the Results; Bono, A. et al. *In Silico* Design of New Dual Inhibitors of SARS-CoV-2 M<sup>PRO</sup> through ligand- and Structure-Based Methods, *International Journal of Molecular Sciences*, 2023, doi: 10.3390/ijms24098377 for the Results).

This work was in part supported by the PJ\_RIC\_FFABR\_2017\_161161 Grant – University of Palermo – and the SiciliAn MicronanOTech Research and Innovation Center “SAMOTHRACE” (MUR, PNRR-M4C2, ECS\_0000022), spoke 3 - University of Palermo - “S2-COMMs-Micro and Nanotechnologies for Smart & Sustainable Communities”.

## Biography and CV of Alessia Bono

July 2021: Graduated in Chemistry and Pharmaceutical Technology with full marks and honors at the University of Palermo with a thesis entitled “Repurposing di derivati benzo[*b*]tiofenici come inibitori selettivi della M<sup>PRO</sup> del SARS-CoV-2” (supervisors Proff. Antonino Lauria and Annamaria Martorana).

October 2021: She started the Ph.D. degree in Molecular and Biomolecular Sciences (University of Palermo), under the supervision of Proff. Antonino Lauria and Annamaria Martorana, focusing his research on the design, synthesis and *in silico* identification of new small molecules with biological activities, specifically as antiviral agents.

October 2023-April 2024: Visiting student at the School of Computing, at the University of Buckingham, under the guidance of Prof. Paul W. Finn, CEO at the Oxford Drug Design (ODD), Oxford Centre For Innovation (OCFI), Oxford.

## List of peer reviewed publications

1. **Bono, A.**; La Monica, G.; Alamia, F.; Lauria, A.; Martorana, A. “A novel *in silico* approach for identifying multi-target JAK/STAT inhibitors as anticancer agents”. Elsevier, Journal of Molecular Graphics and Modelling (ISSN 1873-4243), Volume 135, Article no. 108913, **23 November 2024**, DOI: 10.1016/j.jm gm.2024.108913.
2. La Monica, G.; Alamia, F.; **Bono, A.**; Lauria, A.; Martorana, A. “Scaffold-Hopping Strategies in Aurone Optimization: A Comprehensive Review of Synthetic Procedures and Biological Activities of Nitrogen and Sulfur Analogues”. Multidisciplinary Digital Publishing Institute (MDPI), Molecules (ISSN 1420-3049), Volume 29, Issue 12, Article no. 2831, **13 June 2024**, DOI: 10.3390/molecules29122813.
3. La Monica, G.; **Bono, A.**; Alamia, F.; Lauria, A.; Martorana, A. “Bioisosteric heterocyclic analogues of natural bioactive flavonoids by scaffold-hopping approaches: state-of-the-art and perspectives in medicinal chemistry”. Elsevier, Bioorganic and Medicinal Chemistry (ISSN 0968-0896), Volume 109, Article no. 117791, **9 June 2024**, DOI: 10.1016/j.bmc.2024.117791.
4. La Monica, G.; Pizzolanti, G.; Baimonte, C.; **Bono, A.**; Alamia, F.; Mingoia, F.; Lauria, A.; Martorana, A. “Design and Synthesis of Novel Thieno[3,2-*c*]quinoline Compounds with Antiproliferative Activity on RET-Dependent Medullary Thyroid Cancer Cells”. *American*

*Chemical Society (ACS), ACS Omega* (ISSN 2470-1343), Volume 8, Issue 38, pp. 34640–34649, **11 September 2023**, DOI: 10.1021/acsomeg.3c03578.

5. **Bono, A.**; La Monica, G.; Alamia, F.; Mingoia, F.; Gentile, C.; Peri, D.; Lauria, A.; Martorana, A. “*In Silico* Mixed Ligand/Structure-Based Design of New CDK-1/PARP-1 Dual Inhibitors as Anti-Breast Cancer Agents”. *Multidisciplinary Digital Publishing Institute (MDPI), International Journal of Molecular Sciences* (ISSN 1422-0067), Volume 24, Issue 18, Article no. 13769, **6 September 2023**, DOI: 10.3390/ijms241813769.
6. Mingoia, F.; Di Sano, C.; D'Anna, C.; Fazzari, M.; Minafra, L.; **Bono, A.**; La Monica, G.; Martorana, A.; Almerico, A.M.; Lauria, A. “Synthesis of new antiproliferative 1,3,4-substituted-pyrrolo[3,2-c]quinoline derivatives, biological and *in silico* insights”. *Elsevier, European Journal of Medicinal Chemistry* (ISSN 1768-3254), Volume 258, Article no. 115537, **10 June 2023**; DOI: 10.1016/j.ejmech.2023.115537.
7. **Bono, A.**; Lauria, A.; La Monica, G.; Alamia, F.; Mingoia, F.; Martorana, A. “*In Silico* Design of New Dual Inhibitors of SARS-CoV-2 M<sup>PRO</sup> through ligand- and Structure-Based Methods”. *Multidisciplinary Digital Publishing Institute (MDPI), International Journal of Molecular Sciences* (ISSN 1422-0067), Volume 24, Issue 9, Article no. 8377, **6 May 2023**, DOI: 10.3390/ijms24098377. #
8. Martorana, A.; La Monica, G.; **Bono, A.**; Mannino, S.; Buscemi, S.; Palumbo Piccionello, A.; Gentile, C.; Lauria, A.; Peri, D. “Antiproliferative Activity Predictor: A New Reliable *In Silico* Tool for Drug Response Prediction against NCI60 Panel”. *Multidisciplinary Digital Publishing Institute (MDPI), International Journal of Molecular Sciences* (ISSN 1422-0067), Volume 23, Issue 22, Article no. 14374, **19 November 2022**, DOI: 10.3390/ijms232214374.
9. La Monica, G.\*; **Bono, A.\***; Lauria, A.; Martorana, A. “Targeting SARS-CoV-2 Main Protease for Treatment of COVID-19: Covalent Inhibitors Structure-Activity Relationship Insights and Evolution Perspectives”. *American Chemical Society (ACS), Journal of Medicinal Chemistry* (ISSN 1520-4804), Volume 65, Issue 19, pp. 12500–12534, **28 September 2022**, DOI: 10.1021/acs.jmedchem.2c01005. #

**\*co-first. authorship**

**NOTES:** “PREMIO Ricerc@STEBICEF 2023” as first author – **30 November 2023** – Dipartimento STEBICEF – Università degli Studi di Palermo.

10. Mingoia, F.; Panzeca, G.; Vitale, M.C.; La Monica, G.; **Bono, A.**; Lauria, A.; Martorana, A. “One pot-like regioselective access to 1-aryl-1H-pyrazol-3(2H)-one derivatives and

evaluation of the anticancer activity". ARKIVOC, pp. 191 – 203, **4 April 2022** - DOI: 10.24820/ark.5550190.p011.739.

11. Lauria, A.; La Monica, G.; Terenzi, A.; Mannino, G.; Bonsignore, R.; **Bono, A.**; Almerico, A.M.; Barone, G.; Gentile, C.; Martorana, A. "Antiproliferative Properties and G-Quadruplex-Binding of Symmetrical Naphtho[1,2-b:8,7-b']dithiophene Derivatives". *Multidisciplinary Digital Publishing Institute (MDPI), Molecules (ISSN 1420-3049)*, Volume 26, Issue 14, Article no. 4309, **16 July 2021**, DOI: 10.3390/molecules26144309.
12. La Monica, G.; Lauria, A.; **Bono, A.**; Martorana, A. "Off-target-based design of selective HIV-1 protease inhibitors". *Multidisciplinary Digital Publishing Institute (MDPI), International Journal of Molecular Sciences (ISSN 1422-0067)*, Volume 22, Issue 11, Article no. 6070, **4 June 2021**, DOI: 10.3390/ijms22116070.
13. Lauria, A.; La Monica, G.; **Bono, A.**; Martorana, A. "Quinoline anticancer agents active on DNA and DNA-interacting proteins: From classical to emerging therapeutic targets". *Elsevier, European Journal of Medicinal Chemistry (ISSN 1768-3254)*, Volume 220, Article no. 113555, pp. 1-34, **24 May 2021**; DOI: 10.1016/j.ejmech.2021.113555.

# *Scientific production arisen from this thesis*

### List of conference proceedings

- **Bono, A.**; La Monica, G.; Alamia, F.; Lauria, A.; Martorana, A. "A novel *in silico* approach for identifying multi-target JAK/STAT inhibitors as anticancer agents". Società Chimica Italiana (SCI), Congresso Congiunto delle Sezioni Sicilia e Calabria 2024 (SCISiCa2024). Polo Papardo, Università degli Studi di Messina, **2-3 December 2024**, Messina – Poster number 7 – (Poster Communication, PO07).
- **Bono, A.**; Asaad, A.; Albanese, M.; Cooper, R.I.; Lauria, A.; La Monica, G.; Alamia, A. "Machine learning model for predicting molecular activity using Molecular Descriptors and ElectroShape Descriptors". Merck Young Chemists' Symposium 2024 (MYCS 2024). Hotel Sporting & Hotel Ambasciatori, **12-15 November 2024**, Rimini – (Flash Communication + Poster Communication, FL051).
- Alamia, F.; La Monica, G.; **Bono, A.**; Lauria, A.; Martorana, A. "New small molecules for the Treatment of TNBC: innovative Polypharmacological *In Silico* Approach". Enabling

new alternatives to animal models Workshop. Aten Center, Università degli Studi di Palermo, **04 June 2024**, Palermo – (Oral Communication).

- **Bono, A.**; Martorana, A.; La Monica, G.; Alamia, F.; Mingoia, A.; Lauria, A. “*In silico* design of new dual inhibitors of SARS-CoV-2 M<sup>PRO</sup> through ligand- and structure-based methods”. Ineos Oxford Institute for AMR Research - IOI Early Career Researcher (ECR) Conference 2024, **14 March 2024**, Saïd Business School, Oxford – Poster number 1 – (Poster Communication)
- La Monica, G.; Pizzolanti, G.; Baiamonte, C.; **Bono, A.**; Alamia, F.; Siragusa, G.; Mingoia, F.; Lauria, A.; Martorana, A. “*In silico* lead optimization and synthesis of a new series of thieno[3,2-c]quinoline with antiproliferative activity on RET-dependent medullary thyroid cancer cells”. Società Chimica Italiana - Convegno Congiunto delle Sezioni Sicilia e Calabria 2023 (SCISiCa 2023), **11-12 December 2023**, Palermo – (Oral Communication).
- Alamia, F.; **Bono, A.**; La Monica, G.; Mingoia, A.; Gentile, C.; Peri, D.; Lauria, A.; Martorana, A. “New CDK-1/PARP-1 Dual-Target Inhibitors as a new strategy for the treatment of Triple Negative Breast Cancer”. Società Chimica Italiana - Convegno Congiunto delle Sezioni Sicilia e Calabria 2023 (SCISiCa 2023), **11-12 December 2023**, Palermo, pp. 78 (PC-03) – (Poster Communication)
- **Bono, A.**; La Monica, G.; Martorana, A.; Lauria, A. “New SARS-CoV-2 M<sup>PRO</sup> covalent inhibitors” 2023 – In Bazan G., Fiorica C., Martorana A., Ragusa M.A., Saladino M.L. (eds.). First STeBICeF Young Researcher Workshop, **12 January 2023**. Università degli Studi di Palermo. ISBN: 978-88-942066-1-6 - (Poster Communication, P43).
- La Monica, G.; Martorana, A.; **Bono, A.**; Mannino, S.; Buscemi, S.; Palumbo Piccionello, A.; Gentile, C.; Peri, D.; Lauria, A. “Antiproliferative Activity Predictor: un nuovo tool *in silico* per predire l’attività antiproliferativa nei confronti del pannello NCI60”. Società Chimica Italiana - Convegno Congiunto delle Sezioni Calabria e Sicilia 2022 (SCICaSi 2022), **1-2 December 2022**, Reggio Calabria – (Oral Communication, O24).
- **Bono A.**, La Monica G., Martorana A., Lauria A. “Valutazione *in silico* e sintesi di derivati tieno[3,2-c]chinolinici: potenziali antitumorali RET-targeted nel carcinoma midollare tiroideo”; Società Chimica Italiana, Workshop della Sezione Sicilia 2021 - Catania, pp. 52-53 (P10), **2 December 2021**, Virtual Edition – (Poster Communication).

- Mingoia F., Vitale M.C., La Monica G., **Bono A.**, Lauria A., Martorana A. “New Pyrazolo[1,2-a]benzo[1,2,4]triazine-3-one derivatives as potential SARS-CoV-2 protease modulators”, XXVII CONGRESSO NAZIONALE DELLA SOCIETÀ CHIMICA ITALIANA, BOOK OF ABSTRACTS (ISBN 978-88-94952-24-7), **14-23 September 2021**, Virtual Edition, FAR PO011 – (Poster Communication).