ELSEVIER

KEYNOTE (GREEN)

Are protein–ligand docking programs good enough to predict experimental poses of noncovalent ligands bound to the SARS-CoV-2 main protease?

Ariadna Llop-Peiró [1,†], **Guillem Macip** [1,2,3,†], **Santiago Garcia-Vallvé** [1,*], **Gerard Pujadas** [1,*]

[1] Departament de Bioquímica i Biotecnologia, Universitat Rovira i Virgili, Research group in Cheminformatics & Nutrition, 43007 Tarragona, Catalonia, Spain
[2] CELLEX Research Laboratories, CibeRes (Centro de Investigación Biomédica en Red de Enfermedades Respiratorias. 06/06/0028), Institut d'Investigacions Biomèdiques August Pi i Sunyer (IDIBAPS), 08036 Barcelona, Catalonia, Spain
[3] Pulmonology Department, Hospital Clínic, 08036 Barcelona, Catalonia, Spain

Hundreds of virtual screening (VS) studies have targeted the severe acute respiratory syndrome-coronavirus 2 (SARS-CoV-2) main protease (M-pro) to identify small molecules that inhibit its proteolytic action. Most studies use AutoDock Vina or Glide methodologies [high-throughput VS (HTVS), standard precision (SP), or extra precision (XP)], independently or in a VS workflow. Moreover, the Protein Data Bank (PDB) includes multiple complexes between M-pro and various noncovalent ligands, providing an excellent benchmark for assessing the predictive capabilities of docking programs. Here, we analyze the ability of the three Glide methodologies and AutoDock Vina by using various target structures/preparations to predict the experimental poses of these complexes. Our aims are to optimize target setup and docking methodologies, minimize false positives, and maximize the identification of various chemotypes in a SARS-CoV-2 M-pro noncovalent inhibitor VS campaign.

Keywords: virtual screening; 3CL-Pro; consensus protein–ligand docking; COVID-19; chemotype diversity

Ariadna Llop-Peiró is a PhD student in the Cheminformatics & Nutrition Research Group at Rovira i Virgili University (URV), focusing on discovering new antiviral drugs for potential future pandemics. She previously earned a BSc in biotechnology from URV and an MSc in Cytogenetics and Reproductive Biology from the Autonomous University of Barcelona (UAB). During her time as a student trainee in the Neurobiotechnology group at the Institute for Bioengineering of Catalonia (IBEC), Ariadna contributed to the analysis of the blood–brain barrier. Her work aided in exploring therapeutic transport mechanisms, aiming to establish effective drug administration methods.

Guillem Macip earned his BSc in Biochemistry from UAB in 2018 and an MSc in bioinformatics and biostatistics from Universitat Oberta de Catalunya – Universitat de Barcelona (UOC-UB) in 2019. He completed his PhD with the Cheminformatics & Nutrition Research Group at URV, focusing on identifying inhibitors for the main protease (M-pro) of SARS-CoV-2. Currently, he is a member of the Applied Research in Infectious Respiratory Diseases and Critically Ill Patients Research Group at Fundació de Recerca Clínic Barcelona - Institut d'Investigacions Biomèdiques August Pi i Sunyer (FRCB-IDIBAPS).

Santiago Garcia-Vallvé was awarded a PhD in biochemistry from URV. From 1995 to 2009, his research primarily focused on bioinformatics, sequence analysis, and molecular evolution. Currently a member of the Cheminformatics & Nutrition Research Group at URV, his ongoing research centers on utilizing cheminformatics tools to identify new bioactive compounds for specific targets and pioneering new tools to enhance virtual screening performance. Since 2020, a significant portion of his research has been dedicated to studying SARS-CoV-2, analyzing its mutations, and identifying inhibitors of M-pro.

Gerard Pujadas was awarded a PhD in chemistry from URV in 1998, and then conducted postdoctoral research at Richard Haser's lab (IBCP, Lyon, France). He served as the head of the Department of Biochemistry and Biotechnology at URV from 2013 to 2016. Currently, his research focuses on developing cheminformatic tools to enhance virtual screening performance and the exploration of new antivirals for current and future pandemics.

* *Corresponding authors. Garcia-Vallvé, S. (santi.garcia-vallve@urv.cat), Pujadas, G. (gerard.pujadas@gmail.com).*
† These authors contributed equally to this work.

https://doi.org/10.1016/j.drudis.2024.104137

Introduction

More than 4 years have passed since the outbreak of the global pandemic known as Coronavirus disease 2019 (COVID-2019). Since then, more than 704 million people worldwide have been infected and, of those, over 7 million have died.[p1] The causative agent of COVID-19 is SARS-CoV-2, a positive-sense single-stranded RNA virus from the *Betacoronavirus* genus, which, since December 2019, has been studied in earnest by researchers across the scientific community.[p2] Apart from the spike protein, which is the target of currently approved vaccines, one of the most widely studied viral targets is the SARS-CoV-2 M-pro.[p3] Also known as 3-chymotrypsin-like cysteine protease (3CL-Pro), this protease is involved in cleaving the pp1a and pp1ab polyproteins to produce several nonstructural proteins essential for viral replication and transcription.[p3],[p4] Therefore, it is a key target for the development of antiviral drugs.[p3],[p5]

The activity of M-pro depends on its homodimeric quaternary structure,[p5] and each subunit has 306 residues, constituting three domains (domains I and II are antiparallel six-stranded β-barrel structures, whereas domain III comprises five α-helices; Figure 1). Located between domains I and II, the active site of M-pro differs from other chymotrypsin-like enzymes in that the usual catalytic triad (Ser/Cys-His-Asp/Glu) appears to lack a third component because it comprises the dyad His41 and Cys145 (Figure 1). Thus, it has been suggested that a water molecule fills the role of the Asp/Glu residue in the canonical catalytic triad by mediating interactions between His41 and other conserved residues, such as His164 and Asp187.[p6] Other important parts of the M-pro binding site are the subsites S4, S2, S1, and S1′. According to Zev *et al.*,[p7] S4 comprises the side chains of Met165, Leu167, Pro168, Ala191, and Gln192, and the main chains of Glu166, Arg188, and Thr190; S2 comprises the side chains of His41, Met49, Tyr54, and Asp187, and the main chain of Arg188; S1 comprises the side chains of Phe140, Asn142, Ser144, Cys145, His163, Glu166, and His172, and the main chains of Leu141, Gly143, His164, and Met165; and S1′ comprises the side chains of Thr25, His41, Val42, Asn119, and Cys145 and the main chains of Thr26 and Gly143 (Figure S1 in the supplemental information online).

Numerous studies have attempted to inhibit the activity of M-pro with both covalent and noncovalent inhibitors that bind to the M-pro binding site and engage in a network of intermolecular interactions with various binding site residues.[p3],[p6],[p8],[p9],[p10],[p11] Covalent inhibitors also form a direct bond with the catalytic cysteine residue (Cys145).[p12] Hence, their modes of action are unambiguously different.[p13]

Among the *in silico* approaches in drug discovery, protein–ligand docking stands out as the most popular and widely used. Typically, it is used to eliminate nonsuitable compounds or to validate VS workflows before *in vitro* studies.[p14],[p15] Protein–li-

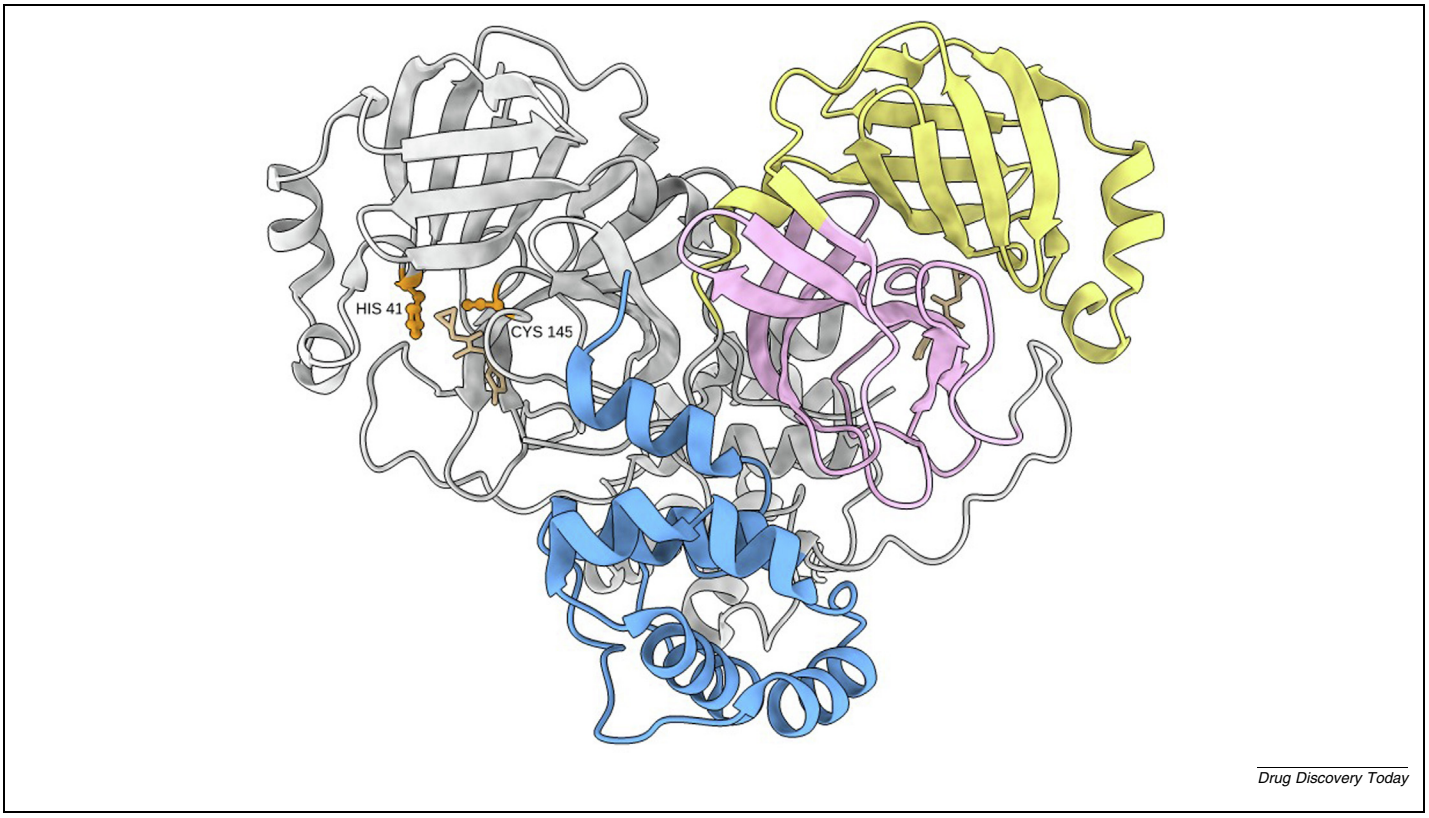


FIGURE 1

Severe acute respiratory syndrome-coronavirus 2 (SARS-CoV-2) main protease (M-pro) homodimeric structure. The color ribbon representation shows the boundaries of each of the three domains in the subunit on the right: domain I is in yellow (residues 8–101), domain II in pink (residues 102–184), and domain III in blue (residues 201–303). The subunit behind (in gray) displays its catalytic dyad (His41 and Cys145) and the noncovalently bound ligand. Figure obtained with the Protein Data Bank (PDB) file 5rgi[p61] and the program ChimeraX.[p62]

gand docking allows billions of compounds to be tested against a protein target at a relatively low computational cost.[p16],[p17],[p18],[p19],[p20] The docking software includes a search algorithm to explore various conformations and orientations of the ligand/compound in a predefined binding site (i.e., docked poses), followed by a scoring function that estimates the binding energy associated with each of the poses generated.[p21] However, as far as M-pro is concerned, protein–ligand docking is controversial because, although it has been successfully used to find some M-pro inhibitors,[p22],[p23],[p24],[p25],[p26],[p27],[p28] the abundance of *in silico* 'findings' has not consistently translated into *in vitro* results.[p29] Therefore, it is imperative to determine whether docking programs struggle to predict the correct binding mode or if the docking score simply fails to properly score the correct/best pose.[p30]

In VS studies, the programs most widely used to identify SARS-CoV-2 M-pro inhibitors are Glide and AutoDock Vina.[p29],[p31] The former has been updated and improved with each Schrödinger release, while AutoDock Vina was not updated to any great extent until 2021, when AutoDock Vina 1.2 was launched.[p32] With its three methods, Glide offers a comprehensive range of speed versus accuracy options: HTVS, SP, and XP. Whereas HTVS and SP docking share the same scoring function, HTVS streamlines the number of intermediate conformations and the thoroughness of the final torsional refinement and sampling.[p33],[p34] Glide XP conducts more extensive sampling compared with SP and initiates with SP sampling before embarking on its anchor-and-grow procedure. XP also incorporates a more sophisticated scoring function designed to filter out false positives (i.e., inactive molecules incorrectly predicted as active) that SP may permit.[p35] AutoDock Vina (often referred to as Vina) is the most widely used open-source program for molecular protein–ligand docking.[p36] It uses a nondeterministic algorithm and, as an open source, has given rise to a variety of other programs (e.g., smina and VirtualFlow).[p37],[p38] It uses an iterated local search global optimizer algorithm, comprising a series of mutation and optimization steps, each of which adheres to the Metropolis criterion, along with a Broyden–Fletcher–Goldfarb–Shanno method for local optimization.[p36] Notably, AutoDock Vina 1.2 was released with new features such as the ability to hydrate the ligand with explicit bridging water molecules before docking.[p32] Interestingly, AutoDock Vina and Glide have been used as protein–ligand docking tools to successfully identify new inhibitors of SARS-CoV-2 M-pro.[p22],[p23],[p28] Other protein–ligand programs that have also succeeded in identifying inhibitors for this target are FRED,[p25],[p39] DOCK3.7,[p27],[p40] and QVina2.[p26],[p41]

The numerous SARS-CoV-2 M-pro structures in the PDB that are complexed with noncovalent ligands have been used by some studies to create a benchmark for analyzing the performance of the most used protein–ligand docking programs (Table S1 in the supplemental information online).[p7],[p31],[p42],[p43],[p44],[p45],[p46] Although these studies have examined the effect of various factors on docking results, such as the presence of water molecules in the PDB files as an integral part of the binding site, the randomization of the initial ligand conformer, the correction of ligand symmetry before root mean square deviation (RMSD) calculation with the experimental pose, and the use of the M-pro homodimeric structure as the docking target, no study has exhaustively considered all these factors simultaneously (Table S1 in the supplemental information online). Therefore, in this review, we evaluate the four docking programs/methodologies most widely used in redocking and cross-docking experiments: Glide HTVS, Glide SP, Glide XP, and AutoDock Vina. We consider all the aforementioned factors simultaneously to determine the extent to which these programs/methodologies can predict the experimental poses of noncovalent ligands bound to SARS-CoV-2 M-pro in the 3D structures of a set of experimental complexes obtained from the PDB. This aspect is crucial, because the performance of protein–ligand docking strongly depends on the PDB coordinates of the target, even for targets such as SARS-CoV-2 M-pro, which exhibit limited flexibility at the catalytic site.[p46] Additionally, we aimed to obtain the same number of docked poses across all our redocking calculations, a precaution that is not always considered in other benchmark studies (Table S1 in the supplemental information online), to avoid bias in our comparisons. Finally, we also consider the role of ligand diversity in evaluating docking success and the potential of consensus protein–ligand docking to reduce the number of false positives in a VS campaign.

Key features and preparation conditions of the noncovalent M-Pro/ligand complexes used in this study

The 47 PDB files containing the SARS-CoV-2 M-pro complexes with noncovalent ligands used in this study were retrieved from the May 8, 2021 update of the PDB and superimposed using Maestro (Schrödinger Release 2023-1). Except for 6m2n, which contained two homodimers (6m2n-1 and 6m2n-2), which were treated as independent entities, the asymmetric unit of all the M-pro complexes contained only one homodimer. Therefore, the total number of complexes included was 48, representing a 40% increase in the number of structures used compared with the largest cross-docking study conducted so far on this target (Table S1 in the supplemental information online).[p31] Nevertheless, since May 2021, the number of noncovalent complexes at the catalytic site of the native SARS-CoV-2 M-pro has increased to 670. However, for practical reasons, especially in the context of the cross-docking analysis, we focused on the 48 complexes originally retrieved.

Figure 2 shows the structural relationships between the noncovalent ligands in the 48 complexes based on their Morgan fingerprints (their 2D structure is presented in Figure 3).[p47] The *scipy.cluster* module in SciPy was used for clustering the ligands according to their structure.[p48] Within the *scipy.cluster.hierarchy* module, hierarchical clustering was performed using the single linkage method. This method computed the hierarchical clustering of the input data and produced a linkage matrix encoding the hierarchical clustering information. Subsequently, the dendrogram function was used to visualize the dendrogram and enable the hierarchical clustering structure to be inspected. In this way, the level at which the dendrogram needed to be cut to obtain clusters was determined. This approach relied on visual inspection of the dendrogram to balance granularity (the number of resulting clusters) and coherence (the separation between clusters). Therefore, with a Tanimoto threshold of 0.61, the 47 ligands

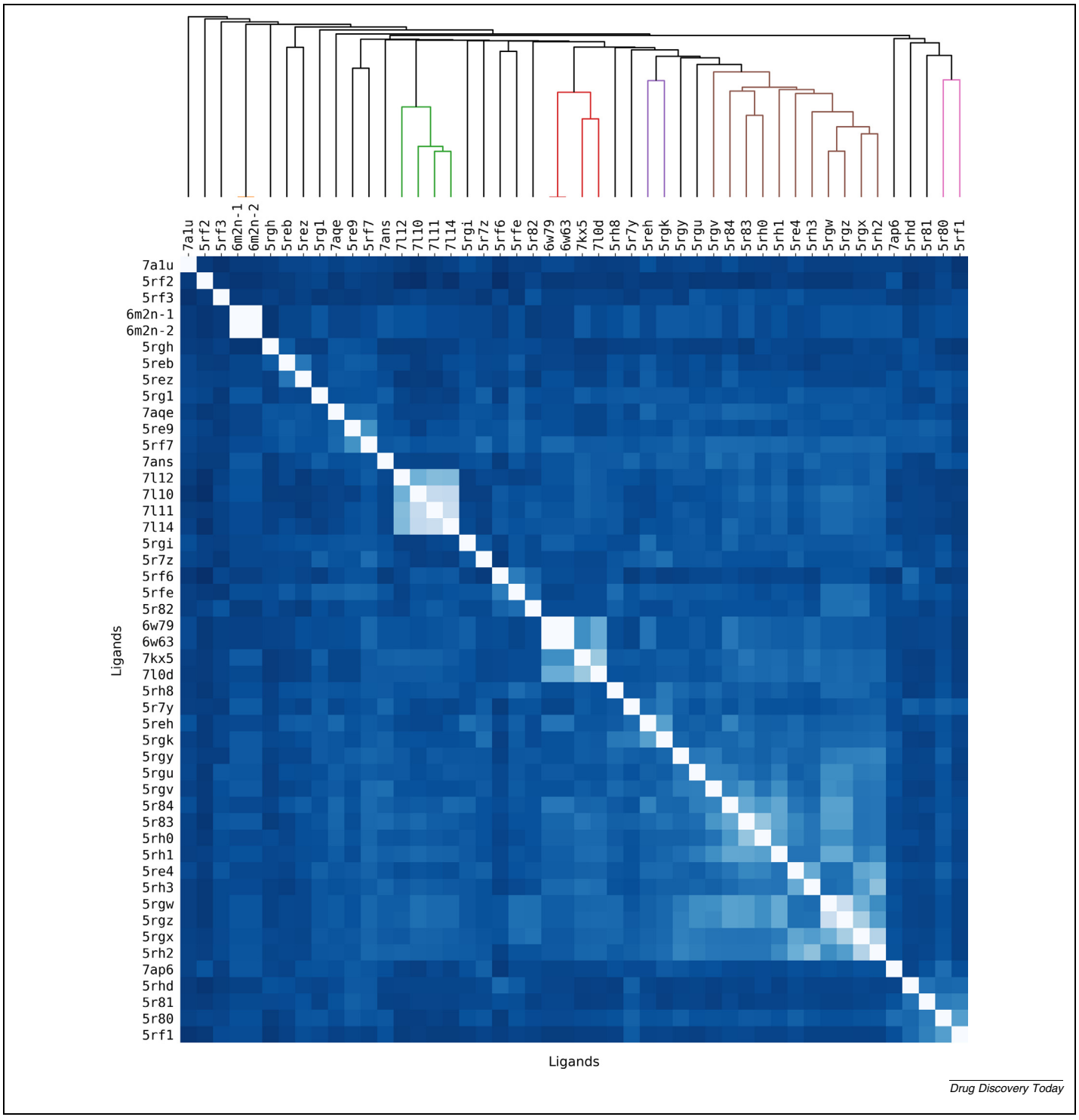
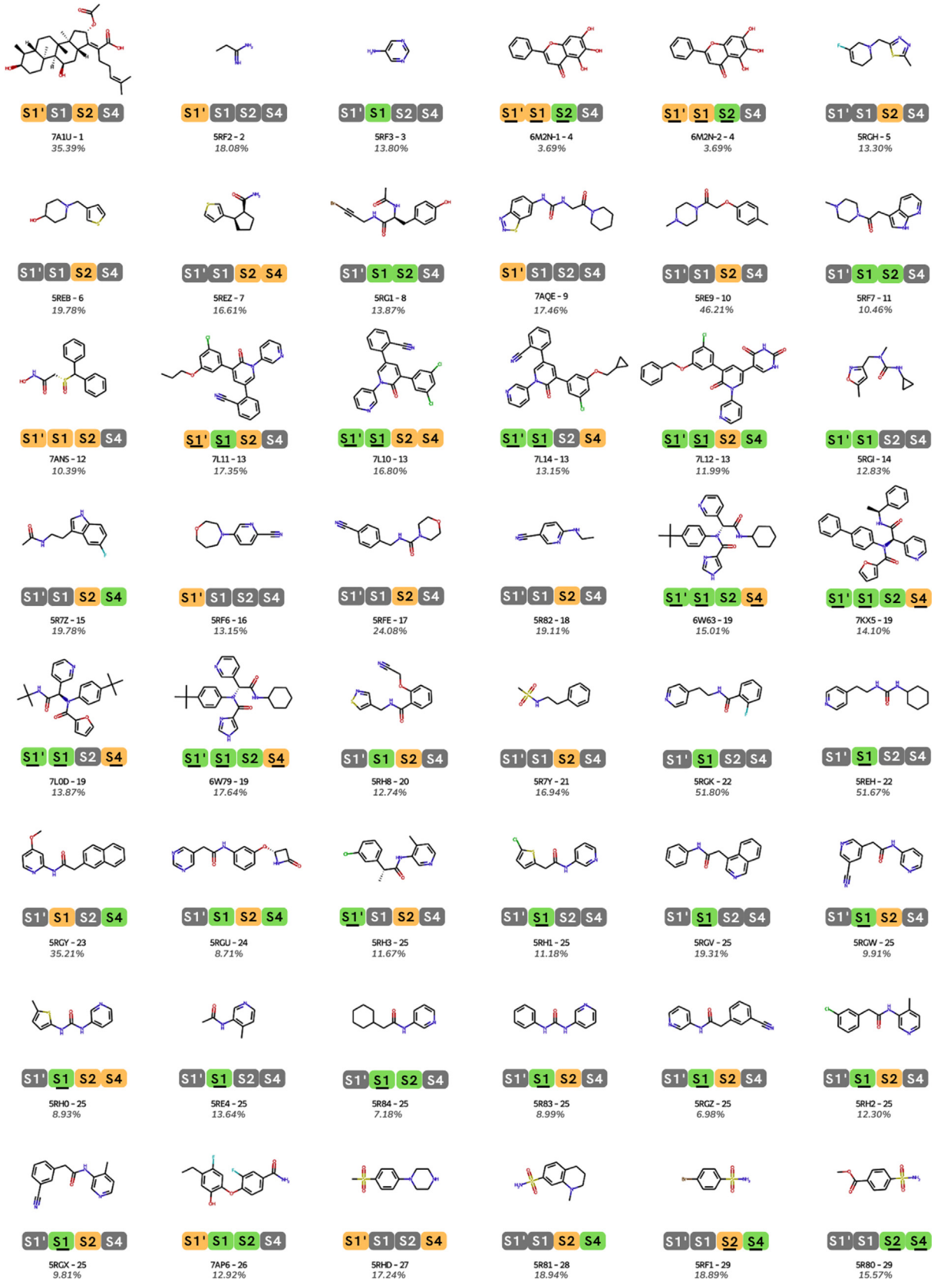


FIGURE 2

Heatmap showing the structural relationships between the noncovalently bound ligands from all 48 complexes analyzed in the current study. Each ligand is labeled with the code of its corresponding Protein Data Bank (PDB) file. Clustering is based on Morgan fingerprints obtained using RDKit (v2022.09.5)[(p47)] with a Tanimoto threshold of 0.61. In the heatmap, similarity is represented by a gradient from blue (indicating completely different compounds) to white (indicating very similar or identical compounds).

were categorized into 29 distinct clusters or chemotypes (Figures 2 and 3). Of these, 24 clusters contained a single ligand (6m2n-1 and 6m2n-2 shared the same ligand in Cluster 4), two clusters contained two ligands (i.e., Clusters 22 and 29), one cluster contained three ligands (i.e., Cluster 19 in which 6w63 and 6w79 shared the same ligand), one cluster contained four ligands (i.e., Cluster 13), and one cluster contained 11 ligands (i.e., Cluster 25). At this point, one might wonder whether using ligands



Drug Discovery Today

from the same cluster would be suitable for our goals and whether it would have been better to use a more diverse set of complexes for our benchmark. Nevertheless, Table S1 in the supplemental information online shows that, of the redocking results obtained by Saar *et al*. with 12 Perampanel derivatives, only 25% of the first docked poses obtained with AutoDock Vina had an RMSD ≤2.0 Å relative to the experimental pose (and 41.7% and 58.3% for Glide XP and Glide SP, respectively).[p44] The same study also showed that the first pose for Glide SP, Glide XP, and AutoDock Vina did not agree with the experimental one for 25% of the Perampanel derivatives (i.e., ligands **6**, **7**, **8**, and **10**).[p44] Therefore, *a priori*, there is no guarantee that closely related compounds will exhibit the same behavior during redocking, which justifies the inclusion of similar chemical structures in our analysis.

Given that the accuracy of the binding site and ligand coordinates in the 48 SARS-CoV-2 M-pro noncovalent complexes is crucial for a proper assessment of the predictive efficacy of the protein–ligand docking programs/methods under evaluation, we used VHELIBS to analyze their fit to the electron density map (EDM).[p49] VHELIBS is designed to simplify the validation of binding site and ligand coordinates for noncrystallographers. Users can specify threshold values for several properties related to the fit of coordinates to electron density (e.g., Real Space R, Real Space Correlation Coefficient, and average occupancy are used by default). VHELIBS then automatically classifies binding sites (i.e., their residues) and ligands as *Good*, *Dubious*, or *Bad* based on these specified limits. Users can also visually inspect the quality of the fit of residues and ligands to the electron density map and reclassify them if needed. In our study, we used the default properties and threshold values in VHELIBS for the initial classification, followed by visual inspection to reassign *Dubious* binding sites and ligands to either the *Good* or *Bad* categories. Subsequently, based on this analysis, the binding site coordinates for nine of the 48 complexes (namely, 5r80, 5r84, 5reb, 5rez, 5rh8, 7a1u, 7ans, 7ap6, and 7aqe) exhibited coordinates categorized as *Bad* by VHELIBS. This could impact the results of protein–ligand docking experiments conducted with these protein structures. Similarly, in five of the 48 complexes (specifically, 5rf2, 5rfe, 5rgi, 5rgk, and 7aqe), VHELIBS identified the coordinates of noncovalent ligands as *Bad*. Therefore, any comparison of docked poses with these ligands might also lead to an underestimation of docking results.

Crystallographic water molecules within 5.0 Å of a ligand atom were meticulously examined to identify those that are spatially equivalent (i.e., within 1.0 Å of another water molecule in a different complex) across the 48 SARS-CoV-2 M-pro complexes (Tables S2 and S3 in the supplemental information online). This analysis makes it possible to assess the impact that solvent molecules have on protein–ligand docking results by comparing three different scenarios in the binding site: (i) *dry* (i.e., no water molecules); (ii) *conserved* (only retaining water molecules conserved in at least 30% of the 48 complexes); and (iii) *water* (all the water molecules originally present in the binding site are retained).

Subsequently, before being used in protein–ligand docking by Glide, all the superimposed complexes were prepared using Protein Preparation Wizard v2023-01.[p50] All the settings were default, except for the following adjustments: (i) the N and C termini were capped; (ii) hydrogens were reintroduced after the original hydrogens had been removed; and (iii) all missing side chains were filled. In the case of AutoDock Vina docking, proteins were prepared and then converted to the specific format (PDBQT) using the *Prepare_receptor.sh* script provided by ADFRsuite v1.0.[p51]

Redocking of noncovalent M-Pro/ligand complexes with known 3D structure

A redocking (or self-docking) experiment evaluates whether a protein–ligand docking program can reproduce a PDB complex by removing the ligand from its corresponding PDB file and subsequently docking it again to the protein. Success is generally measured by whether the RMSD between the docked and experimental poses falls within the 0.0–2.0 Å range.[p7],[p31] Here, redocking was used to assess the performance of the four most common docking programs/methodologies (Glide HTVS, Glide SP, Glide XP, and AutoDock Vina) for replicating the 48 experimental complexes across the three binding site scenarios (i.e., *dry*, *conserved*, and *water*).

Similar to the Firouzi *et al*. benchmark,[p45] special care was taken to eliminate any potential relationship between the input ligand conformation and the experimental pose being predicted to ensure unbiased results. Zajaček *et al*. recently reported that using the experimental pose instead of a randomized one during the redocking of SARS-CoV-2 M-pro noncovalent ligands increased the ability of AutoDock Vina to reproduce the native binding pose of the ligand by 10.2% (they also observed a similar trend for AutoDock 4 and PLANTS).[p43] Here, the input ligand conformation for Glide was randomized by using Maestro to convert the experimental 3D poses of the ligands into 1D SMILES. Subsequently, the SMILES were converted back to 3D using LigPrep v2023-1[p52] with the OPLS4 force field.[p53] During this conversion, the *Do not change ionization* option was selected, and no tautomers were generated. Conversely, for AutoDock Vina, the experimental 3D poses were first converted into 1D using the Open Babel tool[p54] and then a Python code was created to convert chemical structures, represented as SMILES strings, into 3D PDB files using the RDKit library. The code also included the addition of hydrogen atoms. The two sets of chem-

FIGURE 3

2D structure of the noncovalently bound ligands in each of the 48 complexes analyzed in this study, along with the subsites they occupy in the main protease (M-pro) binding site and the percentage of ligand surface area exposed to the solvent. Each ligand is identified by the Protein Data Bank (PDB) code to which it belongs, the cluster in which it is classified, and the subsites it occupies in the M-pro binding site according to Zev *et al*.[p7] Subsites that are completely or partially occupied are colored in green or yellow, respectively, and those occupied by all ligands in the same cluster are underlined. The percentage of ligand surface area exposed to the solvent in the corresponding PDB complex was obtained from Bassani *et al*.[p42] The different ligands are arranged in the order of their clusters from left to right in Figure 2 in the main text.

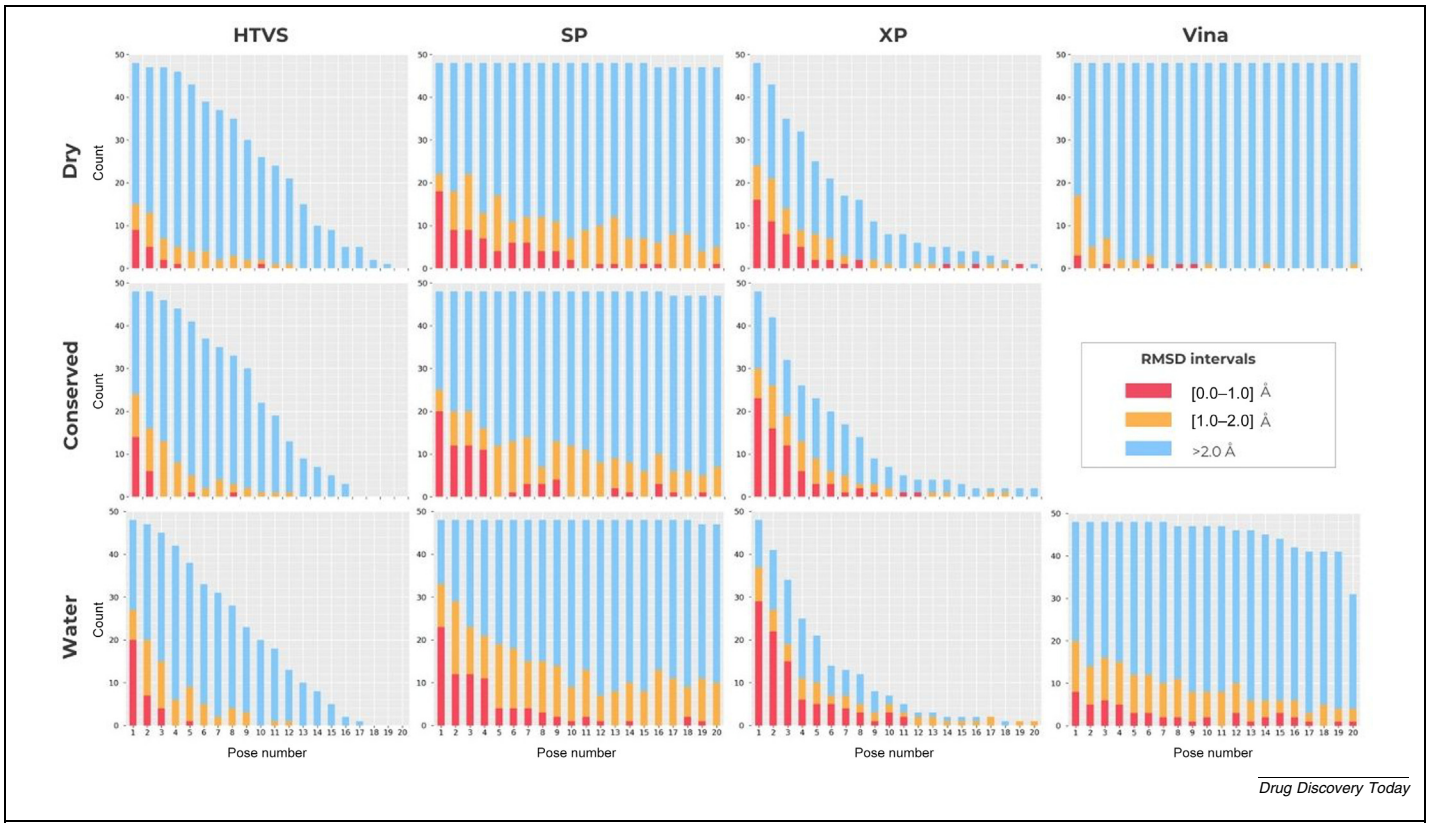


FIGURE 4

Evaluation of redocking success by comparing the root mean square deviation (RMSD) of the first 20 docked poses for each co-crystallized ligand with its experimental pose. The performance of the three protein–ligand docking methodologies available in Glide [i.e., high-throughput VS (HTVS), standard precision (SP), or extra precision (XP)] was assessed in three distinct binding site environments. *Dry* conditions involve the removal of all binding site water molecules before docking. *Conserved* conditions consider only those water molecules in at least 30% of the 48 complexes under study during protein–ligand docking (Table S3 in the supplemental information online). *Water* conditions involve considering all binding site water molecules during protein–ligand docking (Table S2 in the supplemental information online). For AutoDock Vina, redocking success was evaluated in two different binding site environments. *Dry* conditions include removing all binding site water molecules before docking, whereas *water* conditions involve considering explicit water molecules generated by the Meeko Python package during protein–ligand docking. The histogram groups the results into different RMSD ranges.

ical structures obtained through these procedures were visually inspected to confirm that they had been preserved after these transformations. Subsequently, the resulting pairs of conformer sets were used as input for all the corresponding protein–-ligand docking experiments.

The process of grid preparation differs in Glide versus AutoDock Vina. In Glide, the grid for each binding site situation was generated using the *Receptor Grid Generation* panel and the ligand was selected and identified with the *Define Receptor* option. In the case of 6lu7, the grid center corresponded to residue Cys145. By contrast, for AutoDock Vina, the spatial location of the grid was determined by aligning all protein complexes in the Autodock Tools interface. The coordinates of the grid center were (6.57, 1.509, 23.027), and the grid box size was set at $44 \times 50 \times 44$ Å^3. In this context, instead of considering the *conserved* and *water* situations, the Meeko python package was used for the automated addition of explicit water molecules directly to the ligand.[p55]

Subsequently, all the grids generated, along with the 46 ligands in the 48 complexes, were used in redocking experiments conducted with either Glide or AutoDock Vina. All the settings were default with the following exceptions: (i) the number of docked poses retained for each protein–ligand docking was set to 20; (ii) AutoDock Vina exhaustiveness was adjusted to 32; and (iii) the seed number for all AutoDock Vina calculations was consistently set to 831 015 548. Then, the heavy atoms of each of the resulting docked poses were compared with those of their corresponding experimental pose, and the RMSD was calculated to assess the ability of the docking program/methodology to accurately predict ligand binding. Given that symmetry-related misalignments can significantly impact RMSD calculations, ligands bearing symmetric rings, such as those in 5rf3 and 6m2n, were checked manually. The use of the SMARTS method in RMSD calculations within Maestro allows for optimal alignment, which leads to more accurate comparisons and a better understanding of the structural relationships between predicted and experimental poses.[p56] This approach ensures that RMSD values reflect true differences and are not biased by symmetry artifacts.

Figure 4 shows the comparison between each of the 20 docked poses per ligand and their corresponding co-crystallized poses. The resulting RMSD values were then categorized into different ranges to determine the impact of the scoring function on the reliability of the predicted poses. As expected, the first docked

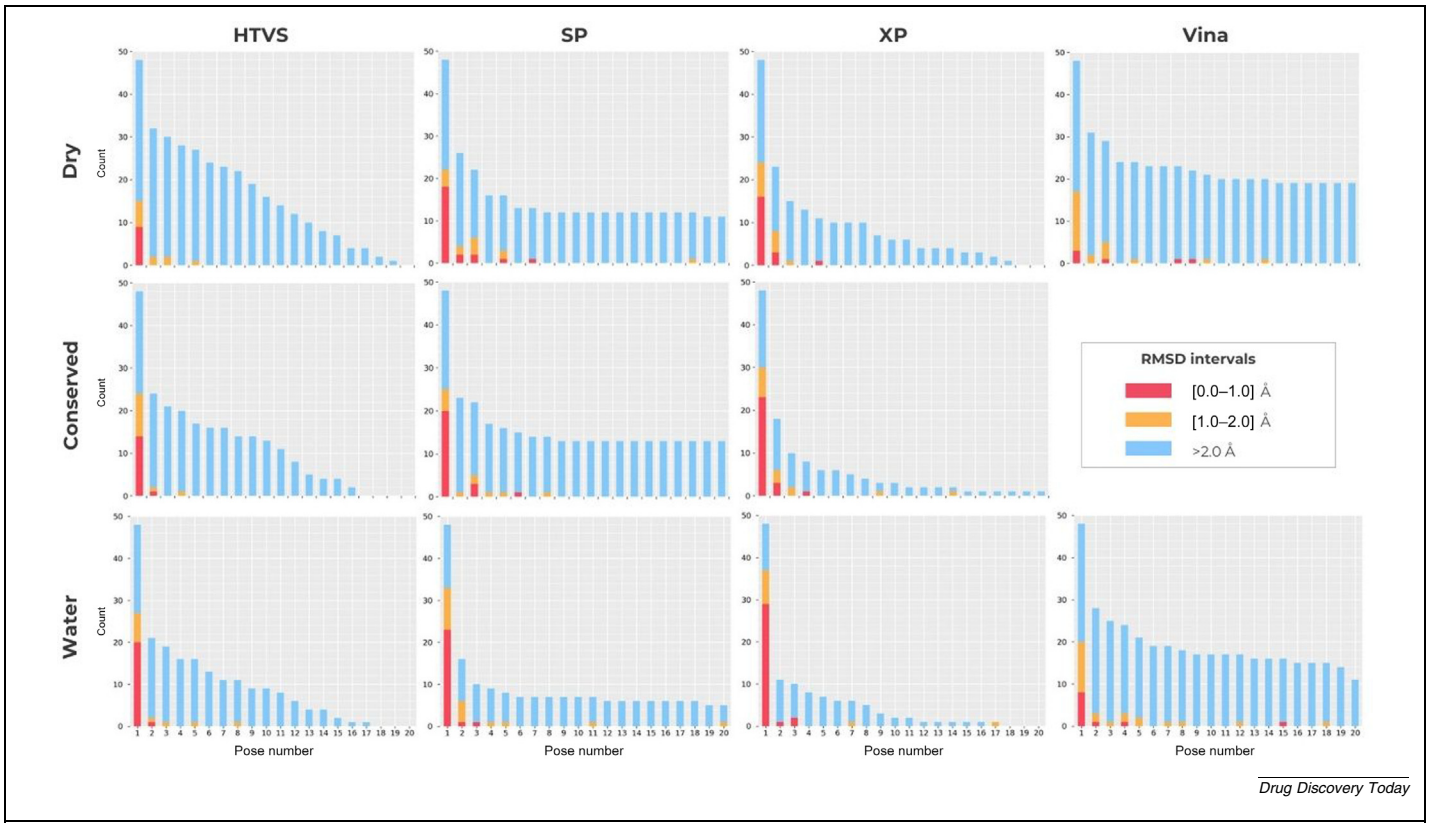


FIGURE 5

Equivalent to Figure 4, but once a redocked pose for a ligand with root mean square deviation (RMSD) ≤2.0 Å was identified, all its lowest-scored poses were excluded from the subsequent comparisons.

pose, generated by various methodologies/programs under different binding site conditions, was the one that most frequently had an RMSD ≤2.0 Å from the experimental pose. In other words, as the pose rank decreased, it became harder for the docked pose to resemble the co-crystallized ligand. Figure 4 further demonstrates that the experimental water molecules in the binding site influenced the conformational sampling algorithm during protein–ligand docking, which increased the number of first poses having an RMSD ≤2.0 Å from the experimental one by *guiding* the docking process.[p7] Consequently, when going from *dry* to *conserved* and *water* conditions, this increase was 15/24/27 for Glide HTVS, 22/25/33 for Glide SP, and 24/30/37 for Glide XP, respectively. These results are in agreement with those obtained by Bassani *et al.*,[p42] who found an increase of 14.1% and 8.9% for Glide SP and Glide XP, respectively, in the number of first poses with an RMSD ≤2.0 Å when all water molecules within 5.0 Å of the co-crystallized ligand during docking were included (Table S1 in the supplemental information online). Similarly, for AutoDock Vina, the incorporation of explicit water molecules enhanced the reliability of the first docked pose in the same RMSD range (increasing from 17 to 20 when comparing the *dry* and *water* scenarios).

Despite the above analysis, crucial questions about the performance of the four programs/methodologies persist, which cannot be addressed by Figure 4 alone. Two of these questions are: (i) can any combination of binding site environments and programs/methodologies reproduce the experimental poses in the 48 complexes with an RMSD ≤2.0 Å?; and (ii) how many docked poses are required to achieve a solution with RMSD ≤2.0 Å from the experimental pose? Figure 5 attempts to answer both these questions by excluding all subsequent low-scoring docked poses after the first instance of an RMSD ≤2.0 Å in comparisons with the experimental pose of a redocked ligand. Remarkably, Figure 5 reveals that, regardless of the combinations of binding site environments and programs/methodologies: (i) the first pose typically tends to have the most RMSD ≤2.0 Å; (ii) no new poses with a RMSD ≤2.0 Å from the experimental pose are identified beyond the 5th ranked one (exceptions are one for Glide HTVS, six for Glide SP, three for Glide XP, and nine for AutoDock Vina); and (iii) some ligands have no docked poses with an RMSD <2.0 Å.

Given that Figure 5 evaluates the redocking performance, the latter result is surprising and, therefore, merits further analysis. To this end, Table 1 provides detailed information on the ligands for which no redocked pose was found with an RMSD ≤2.0 Å. Approximately one-third of redocking failures (i.e., 34.6%) can be attributed to comparing docked poses with experimental poses classified as *Bad* by VHELIBS or to using binding site coordinates also classified as *Bad* by VHELIBS. For the remaining two-thirds of redocking failures (i.e., 65.4%), the key findings from Table 1 are as follows: (i) the number of redocking failures generally decreased for each program/methodology when more water molecules were considered to be part of the binding site (20/16/15 for Glide HTVS, 6/8/4 for Glide SP, 7/5/5 for Glide

TABLE 1

PDB codes for the ligands for which no pose <2.0 Å was found in the redocking experiment using different combinations of binding site environment and docking programs/methodologies[a]

Binding site environment	Docking program/methodology			
	HTVS	**SP**	**XP**	**Vina**
Dry	28/8/**12** **5r7z** 5r80 5r82 5r83 5r84 **5re4** 5re9 5reh 5rfe **5rg1** **5rgh** **5rgk** 5rgv 5rgy **5rgz** 5rh0 **5rh1** 5rh8 **6w63** **6w79** 7a1u 7ap6 7aqe **7kx5** 7l0d **7l10** 7l11 **7l12**	11/8/**3** 5re9 5reh 5rfe 5rgh 5rgk 6w63 7a1u **7ap6** 7aqe **7l0d** **7l10**	14/8/**3** 5r7y 5r80 **5re9** 5reh 5rf1 5rfe **5rgh** **5rgk** 5rh8 6w63 7a1u 7ap6 7aqe 7l0d	19/8/**9** **5r7y** 5r7z 5r82 5r83 **5re9** 5reb **5reh** 5rf1 **5rfe** **5rg1** 5rgh **5rgk** 5rgy 5rgz 5rh0 6m2n-1 **7a1u** **7ap6** **7aqe**
Conserved	21/5/**12**	13/5/**3**	8/5/**3** 5r7y **5re9** 5rfe **5rgh** **5rgk** 5rgy 6w79 7ap6	
	5r7z 5r83 **5re4** 5re9 5rfe **5rg1** **5rgh** **5rgk** 5rgw 5rgy **5rgz** **5rh1** 5rh8 **6w63** **6w79** 7a1u 7aqe **7kx5** 7l0d **7l10** **7l12**	5re9 5reh 5rfe 5rg1 5rgh 5rgk 6w79 7a1u **7ap6** 7aqe 7kx5 **7l0d** **7l10**		
Water	16/1/**12** 5r7y **5r7z** 5r82 **5re4** **5rg1** **5rgh** **5rgk** 5rgv 5rgw **5rgz** **5rh1** **6w63** **6w79** **7kx5** **7l10** **7l12**	5/1/**3** 6w79 **7ap6** 7kx5 **7l0d** **7l10**	6/1/**3** **5re9** 5reh **5rgh** **5rgk** 7kx5 7l0d	14/1/**9** **5r7y** 5r80 **5re9** **5reh** 5rf2 **5rfe** **5rg1** **5rgk** **7a1u** **7ap6** **7aqe** 7kx5 7l0d 7l12

[a]A maximum of 20 poses per ligand were evaluated.

[b]The PDB codes in bold obtained no pose for the given program/methodology, regardless of the binding site environment used. The PDB codes in italics obtained no pose for the given type of binding site environment, regardless of the program/methodology used. The numbers summarize how many ligands were found in the above situations in each cell. The PDB codes for experimental poses and binding sites not validated by VHELIBS are highlighted in red (and underlined for poses).

XP, and 13/7 for AutoDock Vina); (ii) some ligands consistently failed to redock regardless of the number of water molecules in the binding site (11 for Glide HTVS, two for Glide SP and Glide XP, and four for AutoDock Vina); and (iii) some ligands consistently failed to redock regardless of the docking program/methodology used (three, three, and one for *dry*, *conserved*, and *water* binding site environments, respectively). Interestingly, among the noncovalent ligands that mostly failed (even though they and their corresponding binding site had been classified as 'good' by VHELIBS), there were ligands that were small (i.e., 5rgh), medium (i.e., 5re9 and 5reh), and large (i.e., 7kx5 and 7l0d). This suggests that ligand flexibility is not the main factor in these failures (Figure S2 in the supplemental information online). It is also noteworthy that 6w63 and 6w79, which form Cluster 19 with 7kx5 and 7l0d, also failed in a significant number of situations (five and six, respectively) when redocking with any of the Glide methodologies. This suggests that the protein–ligand docking programs/methodologies (especially Glide) performed poorly with this ligand family because they were unable to find a single pose from among the top 20 that resembled the experimental pose. Determining whether this is caused by the conformational search algorithms or the scoring functions requires further investigation and is beyond the scope of the present paper.

Taking all of this into consideration, the primary conclusions drawn from the redocking analysis are as follows: (i) depending on the program/methodology and the conditions used, the experimental pose for 42–90% of noncovalent ligands can be reproduced; (ii) depending on the docking program/methodology and conditions used, pose 1 only agrees with the experimental pose in 31–77% of the ligands (while poses 2–5 and 6–20 agree in 6–27% and 0–10%, respectively); (iii) when water molecules are an integral part of the binding site during docking, it becomes easier for the first pose to correspond to the experimental pose, regardless of the docking program/methodology used; and (iv) if the option requiring less user intervention is selected (i.e., no water present at the active center), then the results for the first pose are best with Glide XP (50%), followed by Glide SP (46%), AutoDock Vina (35%), and Glide HTVS (31%).

In general, the percentage of docked poses with an RMSD $\leq$2.0 Å relative to the experimental pose that we obtained (data in parentheses) was similar to (or higher than) the other comparable benchmarks in Table S1 in the supplemental information online: (i) 42.3–45.5% (46%), 46.2% (50%), and 43.2–63.2% (35%) for the first pose in a *dry* binding site for Glide SP, Glide XP, and AutoDock Vina, respectively; (iii) 55.1–77.3% (77.1%), 53.8% (70.8%), and 52.3–91.5% (60.4%) for all docked poses in a *dry* binding site for Glide SP, Glide XP, and AutoDock Vina, respectively; (iii) 56.4% (68.7%), and 55.1% (77.1%) for the first pose in a binding site with all crystallographic water molecules for Glide SP and Glide XP, respectively; and (iv) 73.1% (91.67%), and 66.7% (87.5%) for all docked poses in a binding site with all crystallographic water molecules for Glide SP and Glide XP, respectively. Although we randomized the conformation of the ligands before performing the docking, our results are comparable to, and in some cases better than, those found in benchmarks that did not randomize.[(p7),(p42)]

It has been suggested that the higher the percentage of ligand surface exposed to the solvent in a crystallographic complex, the harder it is for docking algorithms to reproduce its conformation.[(p42)] Figure S2 in the supplemental information online shows that, although the blue plus signs tend to concentrate in the [0.0–2.0] Å and [0.0–20.0]% areas, a significant portion of the best results for the 48 complexes under study falls in the [2.0–8.0] Å and [0.0–20.0]% area (41.0, 10.3, 12.8, and 17.9% for Glide HTVS, Glide SP, Glide XP, and AutoDock Vina, respec-

tively, in the *dry* situation if the ligands rated as *Bad* by VHELIBS are not considered to calculate the percentage). The same figure shows that, in general, the more rigid a ligand, the better the results during its redocking (red dots). However, there is also a significant portion of the results in the [2.0–8.0] Å and [0–3] rotatable bonds area (27.1%, 10.4%, 12.5%, and 18.8% for Glide HTVS, Glide SP, Glide XP, and AutoDock Vina, respectively, in the *dry* situation).

It has also been suggested that the failure of AutoDock Vina and Glide XP to predict the experimental pose is a consequence of the absence of the S1 subsite.[p46] Although the single noncovalent PDB complex used by these authors to reach their conclusion (i.e., 7ddc) was not in our set of 48 noncovalent complexes, Table 1 (which shows the list of PDB codes for which no pose <2.0 Å was found in the different redocking experiments) and Figure 3 (which shows the subsites occupied by each ligand in the M-pro binding site) demonstrate that this conclusion is only partially true. It is valid for 5rgh and 5re9, which do not bind to the S1 pocket, but not valid for 5rez, 5rtz, 5rf6, 5rh3, 5rhd, and 5r81 (and, to a lesser extent, for 5r82, 5r7y, and 5rf1), which do bind to the S1 pocket. In contrast to Khachatryan *et al.*,[p46] who only considered the first pose in their analysis, we used the first 20 docked poses, which could explain this discrepancy. Figure 5 shows that, for the *dry* binding site situation (equivalent to that considered by Khachatryan *et al.*), only 70.6% and 57.1% of the docked poses with RMSD ≤2.0 Å belonged to the first pose for Glide XP and AutoDock Vina, respectively. Therefore, considering only the first pose might be too restrictive for conclusions to be drawn about the reasons for protein–ligand docking failures in some structures.

Cross docking of noncovalent M-Pro/ligand complexes with known 3D structure

A cross-docking experiment tests whether a protein–ligand docking program can accurately predict the experimental pose of a ligand using a protein structure of the same target but not the one it was originally co-crystallized with. As with redocking, success is measured by whether the RMSD between the docked and experimental poses falls within the 0.0–2.0 Å range. Cross-docking studies are usually similar to the protein–ligand docking step in a VS study (i.e., a target structure is used to dock a library of putative ligands from that target). Therefore, cross docking is a good way to evaluate how protein–ligand docking will behave during a VS.

Here, we determined whether any combination of a docking program/methodology and one PDB structure outperforms other structures in predicting the experimental pose only based on the first docked pose. As demonstrated above, incorporating conserved water molecules or all water molecules into the binding site as an integral part of the receptor structure during grid generation typically made it easier for the initial pose to have an RMSD ≤2.0 Å from the experimental one during redocking (Figure 5). However, the positive impact of these grids during cross docking (or VS) was not clear because the steric hindrance caused by their water molecules (especially crucial in the *water* situation) could result in more false negatives. These false negatives represent the true inhibitors of the SARS-CoV-2 M-pro that might need to displace some of the water molecules from the binding site when they bind to this target. Moreover, for the two protein–ligand docking programs/methodologies that yielded the most favorable outcomes in the redocking experiment, namely Glide SP and Glide XP, the difference in results of the *dry* and *conserved* scenarios for the initial pose was not substantial (increasing from 22 to 25 for Glide SP and from 24 to 30 for Glide XP at RMSD values ≤2.0 Å; Figure 5). Likewise, the hydrated docking methodology recently incorporated into AutoDock Vina is now considered unsuitable for VS, as its authors point out.[p57] Consequently, it was not used in the cross-docking study. Therefore, we assessed the cross-docking ability of Glide HTVS, Glide SP, Glide XP, and AutoDock Vina in a *dry* binding site.

Figures S3–S6 in the supplemental information online depict the cross-docking results obtained for the first pose by Glide HTVS, Glide SP, Glide XP, and AutoDock Vina, respectively (with Figure 6 providing a simplified version of Figure S4 in the supplemental information online). Meanwhile, Table S4 in the supplemental information online summarizes the results of all these cross-docking experiments. These data clearly show that not all SARS-CoV-2 M-pro structures are suitable targets for protein–ligand docking. Particularly interesting is the observation that the 6lu7 structure, commonly used as the target in most VS studies of noncovalent M-pro inhibitors,[p29],[p31] does not appear to be the most suitable for these tasks. It only identifies the ligands of two or three different chemotypes with docked poses that have an RMSD ≤2.0 Å from the corresponding experimental poses for Glide SP or AutoDock Vina. The noteworthy underperformance of 6lu7 with both Glide HTVS or Glide XP is remarkable, considering that the two methodologies have been used in ∼20% of VS studies targeting noncovalent M-pro inhibitors and the PDB structure in 60%.[p29] This suggests that these studies have incorrectly classified a significant number of potentially valuable compounds for COVID-19 treatment as non-active.

Figures S3–S6 and Table S4 in the supplemental information online illustrate that Glide SP appears to be the most effective at detecting diversity in the ligand chemical structure of the VS hits (a critical factor in any VS scenario), followed by Glide XP, Glide HTVS, and AutoDock Vina. This conclusion was drawn because, overall, they were able to identify docked poses with an RMSD ≤2.0 Å from the corresponding experimental poses for 21, 20, 18, and 12 different chemotypes, respectively. Moreover, when this performance was evaluated on a per-target basis rather than overall, Glide SP was confirmed as the best, followed by Glide XP, Glide HTVS, and AutoDock Vina. This is because they accurately docked ligands for at least four different chemotypes in 30, 20, 17, and 9 M-pro structures, respectively.

The top-performing methodology/target PDB combinations in cross-docking were Glide HTVS/5rg1, Glide SP/5rgw, and Glide SP/5rgx, which successfully predicted poses for ligands from eight different clusters (with Clusters 3, 7, 24, 25, and 26 being predicted by all three combinations). Although its results were slightly worse than those of Glide SP, Glide XP performed best with 5rh0 and accurately predicted poses for ligands from seven different clusters (i.e., Clusters 7, 12, 13, 15, 21, 24, and 25). Although the computational time of Glide XP is greater than that of Glide SP, for this specific target at least, this did not increase the accuracy of the results. Interestingly, the perfor-

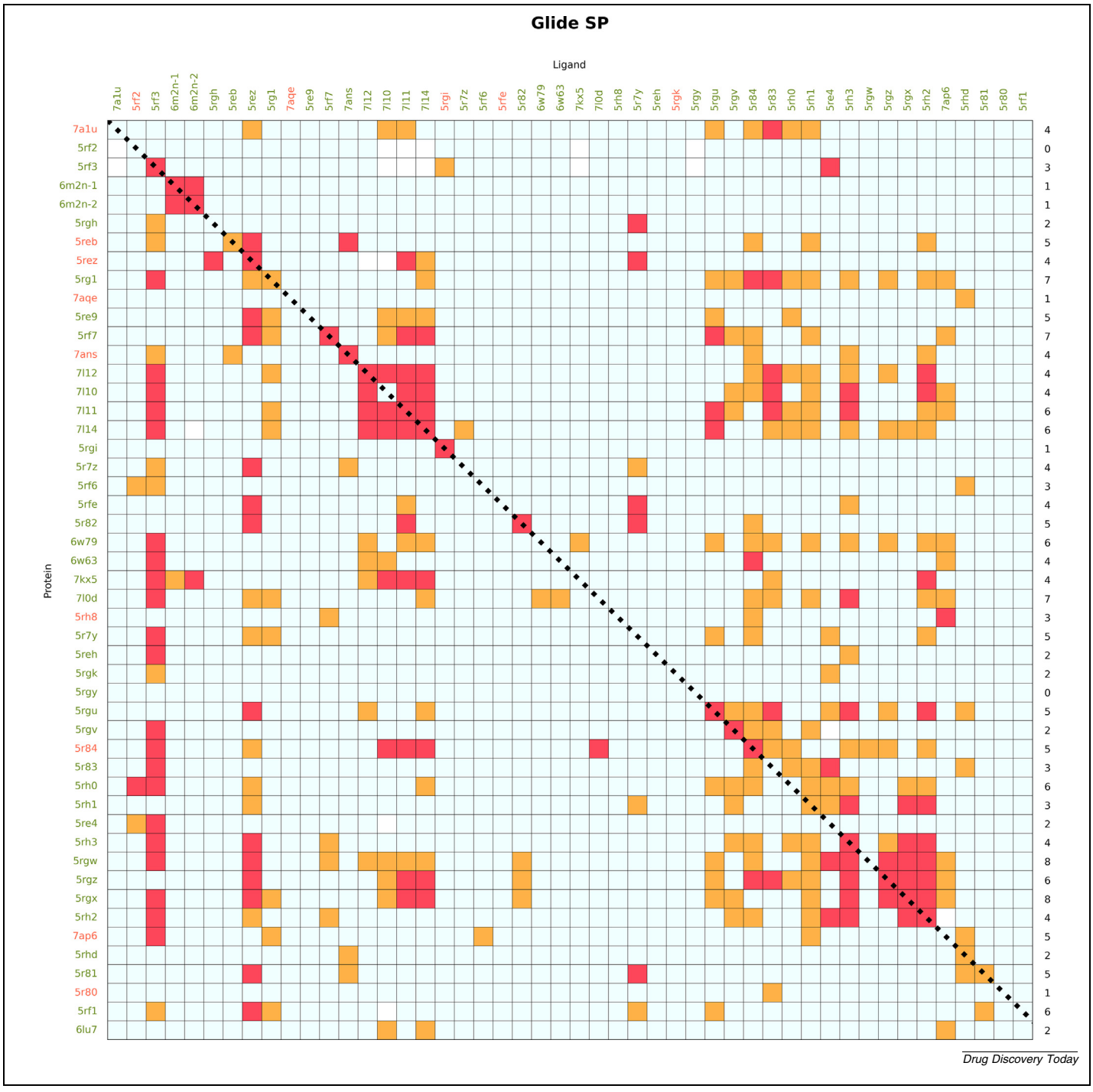


FIGURE 6

Root mean square deviation (RMSD) comparison of the top-ranked docked pose obtained with Glide standard precision (SP) at a *dry* binding site with the corresponding experimental pose. The ligand in each column [identified by the Protein Data Bank (PDB) code of the complex in which it was originally located] was docked to the *dry* binding site of each Severe acute respiratory syndrome-coronavirus 2 (SARS-CoV-2) main protease (M-pro) structure (identified by the corresponding PDB code in each row). The colors of the PDB labels indicate whether the corresponding ligand or binding site coordinates fit (green) or do not fit (red) the electron density map according to VHELIBS.[p49] Thus, the first pose obtained from the docking of a noncovalent ligand and a *dry* binding site was compared with the experimental pose of the corresponding ligand in its complex with the M-pro of SARS-CoV-2, and its RMSD was calculated. The dotted line corresponds to the redocking results and, therefore, is coincident with the results for the first pose in Figures 4 and 5 in the main text in the *dry* binding site. Cross-docking results for 6lu7 (by far the most used SARS-CoV-2 M-pro structure in the literature for docking studies) are also shown in the last row. The same colors used in Figures 4 and 5 in the main text to classify RMSD ranges (red for RMSD $\leq$1.0 Å, orange for RMSD 1.0–2.0 Å, and light blue for RMSD >2.0 Å) were applied to the cells in Figure 6. Ligands and binding sites were sorted in the same order as in Figure 2 in the main text to facilitate the correlation of 2D chemical similarity with cross-docking and redocking results. The numbers at the end of each row represent the number of clusters into which the pairs of docked/experimental poses with RMSD $\leq$2.0 Å were classified ino according to Figure 2 in the main text.

mance of AutoDock Vina (used in ~40% of published VS studies seeking SARS-CoV-2 M-pro inhibitors)[p29] was particularly poor. It correctly identified poses for ligands from a maximum of only five different chemotypes (specifically, docking results for 7l12 and 6w63).

In particular, the binding sites categorized as *Bad* by VHELIBS behaved similarly to the remaining 39 complexes labeled as 'good' (Table S4 in the supplemental information online). In this regard, Glide SP effectively predicted poses for ligands from four or five different clusters, except for 5r80/7aqe and 5rh8, which only did so in one and three different clusters, respectively. Glide XP accurately predicted poses for ligands from four or five different clusters, except for 7aqe, 7ap6, and 5r80/5r84/5reb, which only did so in one, two, and three different clusters, respectively. Likewise, Glide HTVS successfully predicted poses for ligands from four and five different clusters for 5r84 and 7ans, respectively. Furthermore, with this method 5rh8/7aqe and the other five binding sites demonstrated positive results for two and one clusters, respectively. Finally, AutoDock Vina predicted the experimental pose for three and four clusters with 5rez and 5r80, respectively (and 0 clusters for 5reb and two for the remaining six binding sites).

Cross-docking results from other equivalent studies are difficult to compare with ours for various reasons (Table S1 in the supplemental information online). For instance, Llanos *et al.* used three protein–ligand docking programs that differed from those used in our analysis, making a direct comparison with our results inadvisable. Additionally, the data provided in their supplementary materials are not sufficient to corroborate their claim that most of the cross-docking results showed RMSD values <2.0 Å.[p31] Similarly, the analysis by Saar *et al.* with 12 Perampanel derivatives did not use RMSD values to assess cross-docking success.[p44] Instead, they used mean absolute error, Pearson r^2 correlation, or the Kendall τ distance, making direct comparison with our results impossible. Nevertheless, like us, they also concluded that the performance of cross-docking is dependent on the receptor structure. Firouzi *et al.* used the protein coordinates from a set of 17 representative SARS-CoV-2 M-pro structures to dock 48 noncovalent ligands co-crystallized with this enzyme.[p45] All docking poses were then clustered and, for each ligand, the lowest-energy poses from the first and most populated clusters were chosen as representative. By comparing the two representative poses for each of the 48 ligands obtained with AutoDock Vina with the corresponding experimental poses, it was found that, in both clusters, only 16.7% of the poses (i.e., eight ligands) had an RMSD ≤2.0 Å from the experimental one. Unfortunately, this is an overall result, and the authors do not provide details of the individual performance of the 17 protein structures during cross docking, making it impossible to compare their results with ours. Khachatryan *et al.* also used eight representative SARS-CoV-2 M-pro structures to evaluate whether the first docked pose of Glide XP and AutoDock Vina can reproduce the experimental pose at an RMSD ≤2.0 Å during cross docking (Table S1 in the supplemental information online).[p46] Although they provide data on the success rate for each of the eight PDB files and protein–ligand docking programs individually, the lack of distinction between the success of covalently and noncovalently bound ligands makes it impossible to compare their results with ours.

Figures S3–S6 in the supplemental information online also allow us to analyze whether using closely related ligands has been suitable for our benchmark. As previously mentioned, in the redocking results obtained by Saar *et al.* with 12 Perampanel derivatives, only 25% of the first docked poses obtained with AutoDock Vina have an RMSD ≤2.0 Å relative to the experimental pose (and 41.7% and 58.3% for Glide XP and Glide SP, respectively).[p44] Considering our most populated clusters (i.e., clusters 13 and 25 with four and 11 ligands, respectively), we can see that the performance strongly depends on the cluster. For instance, Cluster 13 (comprising four Perampanel derivatives) gave excellent results (93.8% and 100% for Glide SP/XP and AutoDock Vina, respectively). by contrast, results for Cluster 25 were significantly worse (24.8%, 48.8%, 51.2%, and 21.5% for Glide HTVS, Glide SP, Glide XP, and AutoDock Vina, respectively). Therefore, since there is no *a priori* guarantee that closely related compounds will exhibit the same behavior during docking, we can conclude that it is beneficial to include similar chemical structures in this type of benchmark analysis. Although our benchmark only used four Perampanel derivatives (one-third of those used by Saar *et al.*), our cross-docking results with Cluster 13 compared favorably with their redocking results. Remarkably, although all redockings performed by these authors with the ligand co-crystallized at 7l10 (i.e., ligand **10** in their paper) were unsuccessful, we managed to correctly redock it with Glide XP and AutoDock Vina. This is even more noteworthy considering that we randomized the initial conformation of the ligands, which Saar *et al.* did not (Table S1 in the supplemental information online).

How does consensus docking perform during cross docking?

Consensus docking was used during the COVID-19 pandemic to predict new SARS-CoV-2 M-pro inhibitors by means of VS.[p18],[p22],[p58],[p59],[p60] The goal of this strategy is to take advantage of different conformational sampling algorithms to generate hypothetical binding modes. This approach does not rely on a single scoring function to rank the results; instead, it identifies docked poses that are consistently found by different protein–ligand docking programs.[p44]

The results presented above on cross docking within a *dry* binding site form a data set that is perfectly suited to evaluating the effectiveness of consensus docking in this context. This evaluation aims to determine whether the consensus docking approach can distinguish poses that accurately match the bioactive pose and those that do not. This ability is crucial if the occurrence of false positives is to be minimized in a VS that has a protein–ligand docking step. To ensure the reliability of the consensus results, two conformational search algorithms that adopt different approaches need to be compared. To this end, we selected Glide SP (as demonstrated earlier, the most effective Glide methodology in cross-docking) and AutoDock Vina.

Figure 7 displays the protein–ligand pairs in which the RMSD between the first docked pose from Glide SP and AutoDock Vina was ≤2.0 Å (represented by lilac and blue cells). It also shows

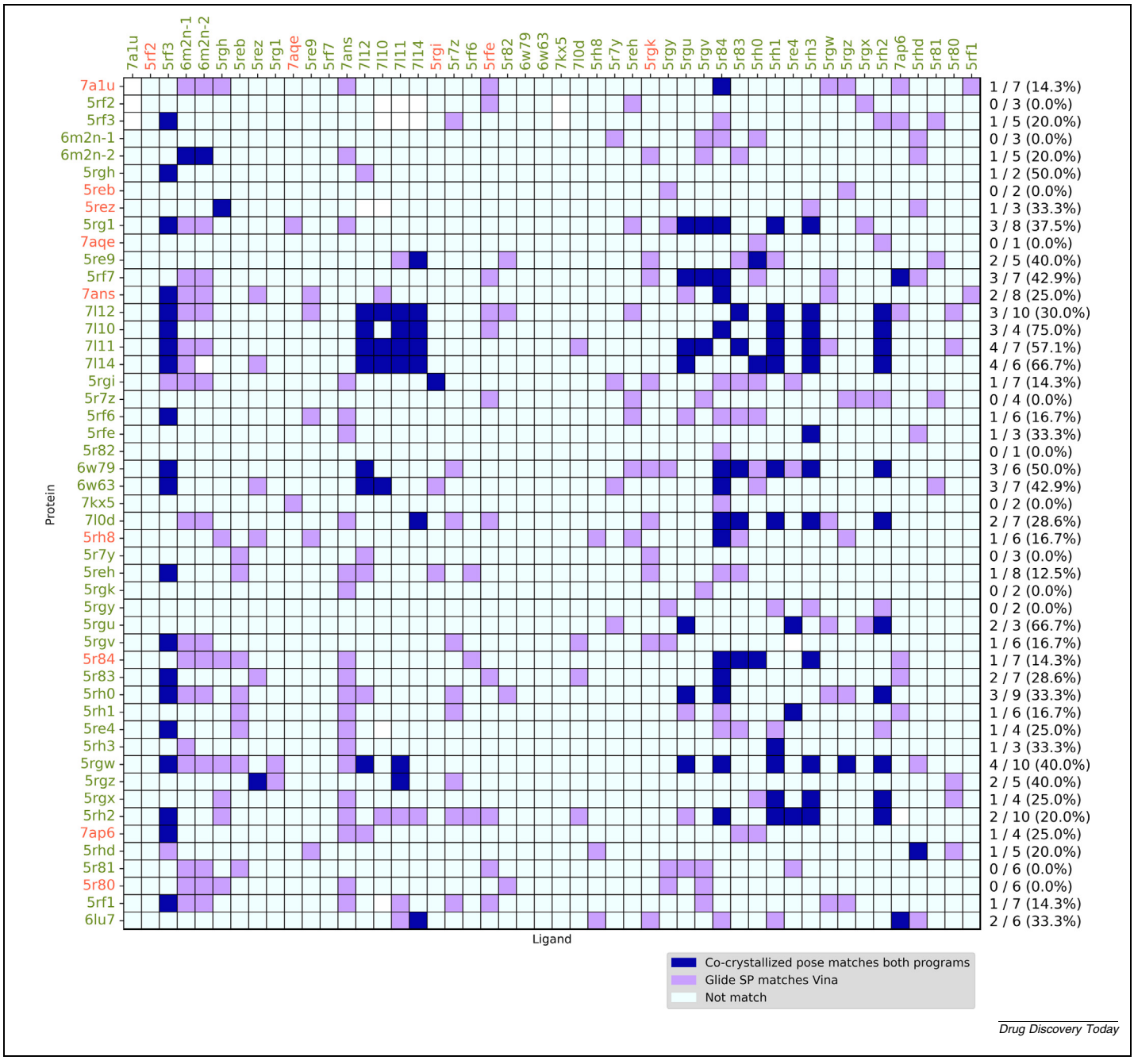


FIGURE 7

Protein–ligand pairs in which the root mean square deviation (RMSD) between the first docked pose from Glide SP and AutoDock Vina within a *dry* binding site was ≤2.0 Å, along with their correspondence to the bioactive pose. Cells in lilac and blue represent the protein–ligand pairs in which the RMSD between the first docked pose from Glide SP and AutoDock Vina was ≤2.0 Å. Blue cells specifically highlight those that were also within ≤2.0 Å of the bioactive pose. The numbers at the end of each row show the count of clusters, which classify the ligands from the blue cells relative to the total number of clusters with ligands from lilac and blue cells (along with the corresponding percentage).

which of these pairs are simultaneously within ≤2.0 Å of the bioactive pose (highlighted in blue). The numbers at the end of each row show the count of clusters, which classify the ligands from the blue cells relative to the total number of clusters with ligands from lilac and blue cells (along with the corresponding percentage). The figure reveals that the success of the consensus docking strategy varied significantly with the protein structure used in protein–ligand docking because the percentage of agreement with the biological pose ranged from 0.0% to 75.0%. For instance, although Glide SP and AutoDock Vina had ligands from six different clusters with RMSD ≤2.0 Å for 5r80 and 5r81, none of them coincided with the bioactive pose. By contrast, the structures that performed best in terms of the number of ligands from different clusters coinciding with the correspond-

ing biological pose were 5rgw, 7l11, and 7l14, each of which have four coincidences.

Reducing the number of false positives in VS is crucial. Therefore, when the consensus docking strategy is used, it is important to prioritize structures that achieve a balance between maximizing the number of coincidences in Glide SP and AutoDock Vina docking results and attaining the highest possible percentage of agreement with the biological pose. The protein structure that best accomplished this balance was 7l10, which had three coincidences (in close proximity to the four of 5rgw, 7l11, and 7l14) and a 75% agreement with the biological pose. Consequently, 7l10 emerges as the optimal protein structure for a VS campaign grounded in a consensus docking strategy with Glide SP and AutoDock Vina because there is likely to be a potentially low number of false positive hits in the consensus results. By contrast, 6lu7 (the target structure in most VS studies of noncovalent M-pro inhibitors)[(p29),(p31)] yielded only two coincidences and a 33% agreement with the biological pose with the same consensus protein–ligand docking strategy.

Concluding remarks

Protein–ligand docking is the primary method in most VS studies designed to discover antivirals to inhibit SARS-CoV-2 M-pro.[(p29),(p43)] Despite the significant computational resources used in this approach, its success rate has been relatively modest. This emphasizes the need for a comprehensive understanding of the strengths and limitations of protein–ligand docking so that it can be optimized before it is used in any research.

One of the limitations revealed by our redocking results is that Glide and AutoDock Vina do not always correctly predict how a ligand binds to its target. This suggests that greater effort needs to be made to not only refine scoring functions, but also enhance pose search algorithms. Another crucial observation is that the choice of the target structure has a clear impact on docking results. Notably, the poor performance of 6lu7, particularly with Glide HTVS and Glide XP (Table S4 in the supplemental information online), resulted in the classification of a significant number of potentially valuable compounds for COVID-19 treatment as false negatives. This is particularly significant in drug repositioning scenarios where the pool of drug candidates is limited, and it is important not to misclassify potentially active compounds that could be beneficial during emergencies, such as the COVID-19 pandemic.

In conclusion, the comprehensive analysis discussed in this review will serve as a guide for best practices in protein–ligand docking in the field of drug discovery, enhancing chemotype diversity and minimizing false positives in VS campaigns. Enrichment validation with actives and decoy sets, which is not covered in this manuscript, should also be considered for target structure selection.

CRediT authorship contribution statement

Ariadna Llop-Peiró: Software, Methodology, Investigation, Formal analysis, Data curation. **Guillem Macip:** Writing – original draft, Investigation, Formal analysis. **Santiago Garcia-Vallvé:** Writing – review & editing, Supervision, Funding acquisition. **Gerard Pujadas:** Writing – review & editing, Supervision, Methodology, Conceptualization.

Acknowledgments

This work is part of the project PID2022-138327OB-I00, financed by the Ministerio de Ciencia e Innovación (MCIN) from the Agencia Estatal de Investigación (AEI) of the Spanish Government (10.13039/501100011033/FEDER, UE). A.L-P. is the recipient of predoctoral grant 2022PMF-INV-14 from the INVESTIGO call funded by the Next Generation EU programme (through the Recovery and Resilience Facility initiative), the Public Service of State Employment (SEPE) of the Spanish Government and the Universitat Rovira i Virgili. All calculations were performed on a workstation equipped with an NVIDIA A6000 GPU, generously provided by Dompé Farmaceutici S.p.A. through the MEDIATE Initiative (https://mediate.exscalate4cov.eu/), which is part of the EXSCALATE4CoV project funded by the EU's H2020-SC1-PHE-CORONAVIRUS-2020 call (Grant No. 101003551). The authors would like to acknowledge the assistance of Júlia Vilalta, Bryan Saldivar-Espinoza, and Aleix Gimeno in preparing some of the figures and tables. This manuscript was edited by the English language service of Universitat Rovira i Virgili.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used ChatGPT to improve language and readability. After using this tool, the manuscript was edited by the English language service of Universitat Rovira i Virgili. Finally, the authors reviewed the content and take full responsibility for the content of the publication.

Appendix A. Supplementary material

Supplementary material to this article can be found online at https://doi.org/10.1016/j.drudis.2024.104137.

References

1. COVID – Coronavirus Statistics – Worldometer. www.worldometers.info/coronavirus/ [Accessed August 12, 2024].
2. Sadeghalvad M et al. Recent developments in SARS-CoV-2 vaccines: a systematic review of the current studies. *Rev Med Virol*. 2023;33(1):e2359.
3. Macip G, Garcia-Segura P, Mestres-Truyol J, Saldivar-Espinoza B, Pujadas G, Garcia-Vallvé S. A review of the current landscape of SARS-CoV-2 main protease inhibitors: have we hit the bullseye yet? *Int J Mol Sci*. 2021;23(1):259.
4. Manelfi C et al. Combining different docking engines and consensus strategies to design and validate optimized virtual screening protocols for the SARS-CoV-2 3CL protease. *Molecules*. 2021;26(4):797.
5. Jin Z et al. Structure of Mpro from SARS-CoV-2 and discovery of its inhibitors. *Nature*. 2020;582(7811):289–293.
6. 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;65:12500–12534.
7. Zev S, Raz K, Schwartz R, Tarabeh R, Gupta PK, Major DT. Benchmarking the ability of common docking programs to correctly reproduce and score binding modes in sars-CoV-2 protease Mpro. *J Chem Inf Model*. 2021;61(6):2957–2966.
8. She Z, Yao Y, Wang C, Li Y, Xiong X, Liu Y. Mpro-targeted anti-SARS-CoV-2 inhibitor-based drugs. *J Chem Res*. 2023;47(4)17475198231184799.
9. Pang X, Xu W, Liu Y, Li H, Chen L. The research progress of SARS-CoV-2 main protease inhibitors from 2020 to 2022. *Eur J Med Chem*. 2023;2573.
10. Janin YL. On the origins of SARS-CoV-2 main protease inhibitors. *RSC Med Chem*. 2023. Published online 2023.
11. Glaser J et al. Hit expansion of a noncovalent SARS-CoV-2 main protease inhibitor. *ACS Pharmacol Transl Sci*. 2022;5(4):255–265.
12. Xiong M, Su H, Zhao W, Xie H, Shao Q, Xu Y. What coronavirus 3C-like protease tells us: from structure, substrate selectivity, to inhibitor design. *Med Res Rev*. 2021;41:965–1998.
13. Awoonor-Williams E, Abu-Saleh AAAA. Covalent and non-covalent binding free energy calculations for peptidomimetic inhibitors of SARS-CoV-2 main protease. *Phys Chem Chem Phys*. 2021;23(11):6746–6757.
14. Gimeno A et al. The light and dark sides of virtual screening: what is there to know? *Int J Mol Sci*. 2019;20(6):1375.
15. Pinzi L, Rastelli G. Molecular docking: shifting paradigms in drug discovery. *Int J Mol Sci*. 2019;20(18):4331.
16. Fink EA et al. Large library docking for novel SARS-CoV-2 main protease non-covalent and covalent inhibitors. *Protein Science*. 2023;32(8):e4712.
17. Luttens A et al. Ultralarge virtual screening identifies SARS-CoV-2 main protease inhibitors with broad-spectrum activity against coronaviruses. *J Am Chem Soc*. 2021;144:2905–2920.
18. Gentile F et al. Automated discovery of noncovalent inhibitors of SARS-CoV-2 main protease by consensus Deep Docking of 40 billion small molecules. *Chem Sci*. 2021;12(48):15960–15974.
19. Schimunek J et al. A community effort in SARS-CoV-2 drug discovery. *Mol Inform*. 2024;43(1)e202300262.
20. Acharya A et al. Supercomputer-based ensemble docking drug discovery pipeline with application to Covid-19. *J Chem Inf Model*. 2020;60(12):5832–5852.
21. Pujadas G et al. Protein-ligand docking: a review of recent advances and future perspectives. *Curr Pharm Anal*. 2008;4(1):1–19.
22. Ghahremanpour MM et al. Identification of 14 known drugs as inhibitors of the main protease of SARS-CoV-2. *ACS Med Chem Lett*. 2020;11(12):2526–2533.
23. Gupta A et al. Structure-based virtual screening and biochemical validation to discover a potential inhibitor of the SARS-CoV-2 main protease. *ACS Omega*. 2020;5(51):33151–33161.
24. Vatansever EC et al. Bepridil is potent against SARS-CoV-2 in vitro. *Proc Natl Acad Sci U S A*. 2021;118(10)e2012201118.
25. Clyde A et al. High-throughput virtual screening and validation of a SARS-CoV-2 Main protease noncovalent inhibitor. *J Chem Inf Model*. 2022;62(1):116–128.
26. Peralta-Moreno MN et al. Shedding light on dark chemical matter: the discovery of a SARS-CoV-2 Mpro main protease inhibitor through intensive virtual screening and in vitro evaluation. *Int J Mol Sci*. 2024;25(11):6119.
27. Luttens A et al. Ultralarge virtual screening identifies SARS-CoV-2 main protease inhibitors with broad-spectrum activity against coronaviruses. *J Am Chem Soc*. 2022;144(7):2905–2920.
28. Unoh Y 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(9):6499–6512.
29. Macip G et al. Haste makes waste: a critical review of docking-based virtual screening in drug repurposing for SARS-CoV-2 main protease (M-pro) inhibition. *Med Res Rev*. 2022;42:744–769.
30. Bellera CL et al. Can drug repurposing strategies be the solution to the COVID-19 crisis? *Expert Opin Drug Discov*. 2021;16(6):605–612.
31. Llanos MA et al. Strengths and weaknesses of docking simulations in the SARS-CoV-2 era: the main protease (Mpro) case study. *J Chem Inf Model*. 2021;61:3758–3770.
32. Eberhardt J, Santos-Martins D, Tillack AF, Forli S. AutoDock Vina 1.2.0: new docking methods, expanded force field, and Python bindings. *J Chem Inf Model*. 2021;61(8):3891–3898.
33. Friesner RA et al. Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J Med Chem*. 2004;47(7):1739–1749.
34. Halgren TA et al. Glide: a new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. *J Med Chem*. 2004;47 (7):1750–1759.
35. Friesner RA et al. Extra precision glide: docking and scoring incorporating a model of hydrophobic enclosure for protein-ligand complexes. *J Med Chem*. 2006;49(21):6177–6196.
36. Trott O, Olson AJ. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J Comput Chem*. 2010;31(2):455–461.
37. Koes DR, Baumgartner MP, Camacho CJ. Lessons learned in empirical scoring with smina from the CSAR 2011 benchmarking exercise. *J Chem Inf Model*. 2013;53(8):1893–1904.
38. Gorgulla C et al. An open-source drug discovery platform enables ultra-large virtual screens. *Nature*. 2020;580(7805):663–668.
39. McGann M. FRED pose prediction and virtual screening accuracy. *J Chem Inf Model*. 2011;51(3):578–596.
40. Coleman RG, Carchia M, Sterling T, Irwin JJ, Shoichet BK. Ligand pose and orientational sampling in molecular docking. *PLoS ONE*. 2013;8(10):e75992.
41. Ding J et al. Vina-GPU 2.0: further accelerating AutoDock Vina and its derivatives with graphics processing units. *J Chem Inf Model*. 2023;63 (7):1982–1998.
42. Bassani D, Pavan M, Bolcato G, Sturlese M, Moro S. Re-exploring the ability of common docking programs to correctly reproduce the binding modes of non-covalent inhibitors of SARS-CoV-2 protease Mpro. *Pharmaceuticals*. 2022;15 (2):180.
43. Zajaček D, Dunárová A, Bucinsky L, Štekláč M. Compromise in docking power of liganded crystal structures of Mpro SARS-CoV-2 surpasses 90% success rate. *J Chem Inf Model*. 2024;64:1628–1643.
44. Saar A, Ghahremanpour MM, Tirado-Rives J, Jorgensen WL. Assessing metadynamics and docking for absolute binding free energy calculations using severe acute respiratory syndrome coronavirus 2 main protease inhibitors. *J Chem Inf Model*. 2023;63(22):7210–7218.
45. Firouzi R, Ashouri M, Karimi-Jafari MH. Structural insights into the substrate-binding site of main protease for the structure-based COVID-19 drug discovery. *Proteins*. 2022;90(5):1090–1101.
46. Khachatryan H et al. Computational evaluation and benchmark study of 342 crystallographic holo-structures of SARS-CoV-2 Mpro enzyme. *Sci Rep*. 2024;14 (1):14255.
47. RDKit: open-source cheminformatics. www.rdkit.org/ [Accessed August 12, 2024].
48. Virtanen P et al. SciPy: Fundamental algorithms for scientific computing in Python. *Nat Methods*. 2020;17:261–272.
49. Cereto-Massagué A et al. The good, the bad and the dubious: VHELIBS, a validation helper for ligands and binding sites. *J Cheminform*. 2013;5(1):36.
50. Protein Preparation Wizard. *Schrödinger Release 2023-1*. New York: Schrödinger; 2023.
51. ADFR software suite downloads. https://ccsb.scripps.edu/adfr/downloads/ [Accessed August 12, 2024].
52. LigPrep. *Schrödinger Release 2023-1*. New York: Schrödinger; 2023.
53. Lu C et al. OPLS4: improving force field accuracy on challenging regimes of chemical space. *J Chem Theory Comput*. 2021;17(7):4291–4300.
54. O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR. Open Babel: an Open chemical toolbox. *J Cheminform*. 2011;3(10):33.
55. GitHub – forlilab/Meeko: Interfacing RDKit and AutoDock. 2023. https://github.com/forlilab/Meeko [Accessed August 12, 2024].
56. Maestro. *Schrödinger Release 2023-1*. New York: Schrödinger; 2023.
57. Hydrated docking: AutoDock Vina 1.2.0 documentation. https://autodock-vina.readthedocs.io/en/latest/docking_hydrated.html [Accessed August 12, 2024].
58. Gimeno A et al. Prediction of novel inhibitors of the main protease (M-pro) of SARS-CoV-2 through consensus docking and drug reposition. *Int J Mol Sci*. 2020;21(11):3793.
59. Ochoa R, Palacio-Rodriguez K, Clemente CM, Adler NS. dockECR: open consensus docking and ranking protocol for virtual screening of small molecules. *J Mol Graph Model*. 2021;109, 108023.
60. Mateev E, Georgieva M, Zlatkov A. In silico identification of novel SARS-CoV-2 main protease and nonstructural protein 13 (nsp13) inhibitors through consensus docking and free binding energy calculations. *Comb Chem High Throughput Screen*. 2023;26(6):1242–1250.
61. Douangamath A et al. Crystallographic and electrophilic fragment screening of the SARS-CoV-2 main protease. *Nat Commun*. 2020;11(1):5047.
62. Pettersen EF et al. UCSF ChimeraX: structure visualization for researchers, educators, and developers. *Protein Science*. 2021;30(1):70–82.