

Dock-able linear and homodetic di, tri, tetra and pentapeptide library from canonical amino acids: SARS-CoV-2 Mpro as a case study.

Sarfraz Ahmad¹, Muhammad Usman Mirza^{1*} John. F. Trant^{1*},

¹Department of Chemistry and Biochemistry, University of Windsor, Windsor N9B 3P4, ON, Canada.

***Corresponding authors:** mumirza@uwindosr.ca; jtrant@uwindosr.ca

Abstract

Peptide-based therapeutics are increasingly coming to the forefront of biomedicine with their promise of high specificity and low toxicity. Although noncanonical residues can always be used, employing only the natural 20 residues restricts the chemical space to a finite dimension allowing for comprehensive *in silico* screening. Towards this goal, the dataset comprising all possible di, tri, and tetrapeptide combinations of the canonical residues has been previously reported. However, with increasing computational power, the comprehensive set of pentapeptides are now also feasible for screening as are the comprehensive set of cyclic peptides comprising four or five residues. Here, we provide both the complete and prefiltered libraries of all di, tri, tetra, and pentapeptide sequences from 20 canonical amino acids and their homodetic (N-to-C-terminal) cyclic homologues. The libraries in the FASTA, SMILES, and SDF-3D formats can be readily used for screening against protein targets. Access to this dataset will accelerate small peptide screening workflows and encourage their use in drug discovery campaigns. As a case study, the developed library was screened against SARS-CoV-2 Mpro to identify potential small peptide inhibitors.

Keywords: Dipeptide, tripeptides, tetrapeptides, pentapeptides, N-to-C-terminal, cyclicpeptides, peptide library.

Introduction

With few exceptions including insulin, therapeutic peptides have only recently become increasingly prevalent. Approximately 10% of the drugs approved between 2009 and 2013 were peptide-derived, and most of these compounds are first-in-class drugs, *e.g.*, bivalirudin, desmopressin, octreotide, *etc.* [1]. As of this writing, more than 80 unique peptide drugs have been approved worldwide, while over 150 are in development [2, 3].

Most drugs work by interfering, strengthening, or modifying protein-ligand interactions. This implies ligand in the broadest possible definition: small molecules, ions, proteins, nucleic acids, lipid membranes, or peptides. In fact, over 40% of all intracellular interactions involve peptide-protein interactions, and many more involve protein-protein interactions. In both cases, peptides and peptidomimetics are ideal interfering agents [4]. Cyclic peptides show exceptional promise in specifically targeting “undruggable” “blank wall” surfaces of proteins involved in protein-protein interactions. A second factor underlying their utility is their perceived niche between small molecules and biologic therapeutics [5].

Biologics comprise a broad range of polysaccharides, proteins, nucleic acids or their conjugates, *e.g.*, vaccines, blood components, somatic cells, and recombinant proteins. They are usually produced in living organisms (animals, microorganisms, or even humans) and then processed *ex vivo*. They often offer treatment options for disorders with no available treatments [5, 6]. However, biologics are often complex mixtures, difficult to characterize. Furthermore, they are generally high-cost, low stability, heat sensitive, susceptible to microbial contamination, and immunogenic [7, 8] (**Figure 1**). Due to their low membrane permeability, susceptibility to digestion, and first-pass metabolism, they are also usually administered parenterally, resulting in challenges for patient compliance [9].

In contrast, small molecules (usually < 1000 Da) offer an enormous amount of chemical diversity to interact with their targets. Inherent advantages typically associated with small molecules include low cost, high stability, good bioavailability, and easy administration [5]. They often effectively bind to small binding pockets in target proteins but have limited application in blocking protein-protein interactions due to their small surface area [10]. However, this small size means they can often interact with other proteins, and many exhibit off-target effects and toxicity. Their non-endogenous metabolites can also induce undesired physiological responses [11].

Small peptides lie between both classes of therapeutics (**Figure 1**). They are often better tolerated, more specific than small molecules, lower cost, and more stable than biologics. They share the single characterizable entity benefit of small molecules but can present specificity similar to biologics [12, 13]. Tetra and pentapeptides offer an optimum size to occupy catalytic or molecular recognition sites often occupied by small molecule metabolites or the termini of endogenous peptides or proteins [14].

Larger peptides can imitate proteins to block surfaces. Together these opportunities have driven the rise of peptides [3, 15-17].


 Superior	Small molecules	Biologics	Biologics	Small molecules	Small molecules
	Peptides	Peptides	Peptides	Peptides	Peptides
	Biologics	Small molecules	Small molecules	Biologics	Biologics
	Cost-effective	Less toxicity	Specific	Less immunogenic	Membrane permeability

Figure 1: Comparison of small molecules, peptides, and biologics for their benefits and drawbacks as drugs.

One advantage of working with peptides is that they occupy a very large, but finite chemical space, with the caveat that one considers only the canonical residues. This means that it can be comprehensively explored *in vitro* or, theoretically, *in vivo*. These are still largely cost-prohibitive, but a preliminary comprehensive, *in silico* screen is becoming feasible.

Many docking algorithms are available for protein-protein and peptide-protein docking that can accommodate the conformational restrictions on the inherently flexible nature of peptides and specific amino acids. Various benchmark studies have been reported assessing the precision of these algorithms, and they are constantly improving [18, 19]. Despite this progress, high throughput screens of small peptides in computer-aided drug discovery (**CADD**) campaigns are still not frequently performed, possibly due to the large number of calculations required, difficulties in handling the conformational complexities of peptides, and a more computationally expensive workflow compared to small molecules. But a primary limitation is that comprehensive machine-readable peptide libraries of dockable peptides with all possible sequences are unavailable.

To address these challenges, we have developed, and here make available, the comprehensive library of all possible di, tri, tetra, and pentapeptides sequences from 20 canonical amino acids. We have also provided a library of N-to-C-terminal cyclic analogues of these peptides. To demonstrate the potential of this tool, as part of our broader effort in CADD, we provide a case study screening this tool against the target-du-jour, the main protease of SARS-CoV-2 (SARS-CoV-2 Mpro or 3CLpro), a cysteine protease [20].

Although the novel coronavirus has a unique Mpro, it is still very similar to that of the original SARS-CoV-1 and was a well-understood enzyme and a well-investigated drug target [21]. This, of course, accelerated with the emergence of SARS-CoV-2 in late 2019 and early 2020, and the crystal structure of the SARS-CoV2 was solved extremely quickly and by multiple efforts nearly simultaneously [22-24] as part of what is the most accelerated drug discovery effort in human history. Mpro cleaves

polypeptides after a glutamine residue, making it a promising drug target because no human proteases share this substrate specificity, meaning off-target effects can be expected to be minimal [24, 25]. The active site of Mpro consists of a catalytic dyad of Cys145 and His41. Other anchoring residues important in ligand interactions in cocrystallized structures include Thr25, Thr26, Asn142, Gly143, Ser144, Met165, Glu166, and Gln189 [26].

Inhibiting this protein would prevent the virus from processing the polypeptide properly and would consequently prevent capsid formation and release from the infected cell; this would halt infection [27]; however, despite very significant effort, there are no drugs in advanced clinical trials that are highly effective, and despite the efforts from many academic groups suggesting drug repurposing might prove helpful, it would be incredibly naïve to consider that all relevant extant drugs were not investigated immediately by major players within weeks and months of the emergence of the disease. Furthermore, the various efforts, 61 independent reports evaluated by Macip and coworkers, do not agree or converge around a common series [28]: repurposed drugs have been thoroughly screened *in vitro* [29]. Although some are initially promising [30-33], none are effective, including those that were developed to target the main proteases of other related viruses: the active site is significantly different, even from closely related SARS-CoV-1 [34], and the entire protein is highly dynamic for a globular enzyme that makes small molecule drugs difficult to design. New chemical material is required.

The protein's active site needs to be dynamic as it must cleave the polypeptide at various different points along its sequence. The natural ligand is a peptide. One molecule that has shown some activity is a peptidomimetic arising from a high-throughput screen of Ugi reaction products [35]. Although no natural peptide is likely to be a useful drug in itself [36], they might prove to be useful starting points for structure-activity relationships through both incorporating unnatural residues and modifying the backbone to increase the stability of the peptide (both physiologically and to avoid peptidase activity in the binding site) and affinity for the target [37]. But for this to be viable, we require a method to identify promising peptides to be validated *in vitro* as potential inhibitors. This makes SARS-CoV-2 Mpro a potentially useful subject for our peptide screening library to be deployed against.

2. Methods

2.1 Preparation of peptide library

Single letter codes of all 20 canonical amino acids (A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, and Y) were used to make all possible sequences of di-, tri-, tetra-, and pentapeptides based on the relation 20^n ($n = 2, 3, 4, \text{ and } 5$) using MS Excel. These peptides were either filtered for their physico-chemical properties (solubility, stability, and ease of synthesis, see next section) based on their amino acid sequence or linked via N-to-C-terminal amide linkage to get cyclicpeptides.

2.1.1 Peptide filtration based on physico-chemical properties

The interplay between the constituent amino acids of a peptide is responsible for its solubility, stability, and ease of synthesis. A set of seven filters was carefully selected based on stability, solubility, and ease of synthesis. Although there are many complex tools that could be used to conduct these filtrations, the digital nature of the peptide sequences (consisting of only 20 values), makes them extremely easy to manipulate in Excel, which also lowers the barrier to use for others. Excel also makes it simple to modify the filters we applied, either expanding them should the user believe we were insufficiently restrictive in our selection process or reducing them should the user believe we were overly zealous in our elimination of sequences. This can be accomplished through simple manipulation of the provided formulae. However, in the case of a reduction of the filters, we caution that the user will need to optimize additional peptide conformations. In all cases, we have provided examples of the process in the supporting information that can be applied to the full dataset.

Filter 1: Peptides that can form gels in an aqueous solution

Peptides having a very high ratio (>75%) of amino acids that can form intermolecular hydrogen bonds (D, E, H, K, N, Q, R, S, T or Y) reliably result in gel formation in water, rendering them water-insoluble. We removed these peptides through simply eliminating the entries in the Excel file. This was accomplished by copying the column containing all sequences (Column A) to Column B. Using the “replace” function on Column B only, substitute “X” for all of D, E, H, K, N, Q, R, S, T, and Y. To remove peptides that are comprised of >75% of above stated residues, cells in column B containing >75% X were deleted. This is achieved by using replace option of Excel for the column: for a pentapeptide, put *X*X*X*X* in the “find” field, leaving the “replace” field blank. All cells containing four or five Xs (4 Xs is 80%) will become blank. Sort the entire worksheet by Column B. This will put the blank cells at the top. Remove all rows that have no entry in column B. Column B can then be deleted. This removes all of these peptides from the list, and this new list can then be applied to the next filter.

Filter 2: Peptides with 50% or more hydrophobic residues are insoluble or partly soluble in water

We can employ an analogous method to remove entries with >50% hydrophobic residues (A, I, L, M, F, W, Y, and V). Again, duplicate column A to column B, convert the hydrophobic residues to X, and replace *X*X*X* with a blank for pentapeptides with blank cell. Once the empty rows are sorted and deleted, these sequences are removed.

Filter 3: Multiple cystines and methionines in a peptide are prone to oxidation. Multiple cystines can also form a disulfide bridge.

This filter should be run at the discretion of the user. As the peptides examined here are so small, intramolecular disulfide bridges are unlikely, and free cysteines could be more likely to form intermolecular bridges. So any *in vitro* test of these peptides might provide data about complex

mixtures of oligomers rather than the planned peptide. Although these peptides might be valid, the complexities they could introduce into both synthesis and screening merits, in our opinion, their exclusion. The filter was applied as above by replacing any strings containing *C*C* and *M*M* with blank cells to remove entries having two or more cystines or methionines.

Filter 4: Multiple prolines or serines in a small peptide can readily cause deletions during synthesis. Multiple prolines can also experience cis/trans isomerization.

Again, this filter may or may not be appropriate for a given application. We are employing it to ensure a higher hit-rate for our peptides in any *in vitro* test. To employ this filter, replace *S*S* and *P*P* with blank cells to remove entries having two or more serine or proline residues.

Filter 5: Aspartic acid next to glycine, serine, or proline can readily undergo hydrolytic peptide cleavage in an acidic solution.

Again, this filter may not always be appropriate, and with sufficient care in synthesis, this complication can be evaded. However, for a simple screen, this challenge could introduce challenges. To employ this filter, removing aspartic acid glycine pairs, remove *GD* and *DG* with blank cells. Similarly, use *PD* and *DP* to remove aspartic acid and proline pairs and *SD* and *DS* to remove aspartic acid serine pairs.

Filter 6: N-terminal glutamine can form pyroglutamate in an acidic solution.

This could even be desirable under specific conditions. But it can be simpler to simply exclude these examples. To execute this filter, put the FASTA data in column A and put : in all cells of column B. Use formula “=B1&A1” to get : at the start of every entry (left-hand side, *i.e.*, N terminal). Paste the formula-generated column to column C as text to avoid the formula. Remove columns A and B. Put :Q* in find and leave replace field blank to remove sequences starting with Q. This same syntax can be used to remove any other problematic N-terminal sequence.

Filter 7: Protecting group on N-terminal asparagine are difficult to remove.

This filter was employed simply to ease synthesis. Again, this is solvable, but does require specialist peptide synthesis knowledge and may not be appropriate for all groups looking to screen and prepare peptides. This filter can be avoided. It was implemented using the protocol used in Filter 6, but employing :N*.

Note: These filters were only performed on pentapeptides. Filtered pentapeptides and all tri and tetrapeptides were placed in text files in FASTA format along with their FASTA identifiers. The files were loaded in MarvinView [38], and the structures were saved as SMILES (simplified molecular-input line-entry system) [39] and SDF (structure-data file) files along with their FASTA [40] identifiers for later use in docking.

2.1.2 Preparation of cyclic peptides

All possible combinations of di-, tri-, tetra- and pentapeptides were used without any prefiltration to generate head-to-tail cyclic amide-linked peptides. To generate this library, the peptides in FASTA format were placed in Excel in column A. Column B was populated with X (the FASTA notation for a general amino acid with an arbitrary sidechain). The formula “=B1&A1&B1” was applied in column C to get X at the start and end of each peptide (*i.e.*, XpeptideX). These peptides were copied in a .txt file, and the file opened in MarvinView. The structures were saved in SMILES format. The resulting SMILES file was opened in notepad++, and the following replacements were made:

For N terminal: ([H])C(=O)[C@]([H])([*])N([H]) to 9, and (C(=O)[C@]([H])([*])N([H])[H]) to 9.

For C Terminal: [H]OC(=O)[C@]([H])([*])N([H])C to C9, and C(=O)N([H])[C@@]([H])([*])C(=O)O[H] to C9(=O).

The FASTA identifiers were placed next to the SMILES separated by a space. The resulting SMILES were filtered for duplicates using Open Babel [41, 42] through the option “remove duplicates with descriptor” InChI (the IUPAC international chemical identifier). After the removal of the duplicate entries, the structures were exporting as SMILES files. The SMILES files were loaded in MarvinView and the structures were saved as an SDF file for docking.

2.2 Virtual screening

2.2.1 Ligand preparation

The ligands in the SDF file were imported into Maestro, and the 3D structures of the compounds were batch prepared using the LigPrep utility of Maestro (Schrödinger Release 2020-4: Maestro, LigPrep, Schrödinger, LLC, New York, NY, 2020) [43]. The possible tautomers of the compounds were produced using Epik [44] at the target pH of 7±2. Only one structure was produced for each ligand.

2.2.2 Protein preparation

The crystal structure of SARS-CoV-2 Mpro protease was downloaded from the Protein Data Bank (PDB ID 6XR3, resolution 1.45 Å) [45]. The Protein Preparation Wizard of the Maestro molecular modeling software (Schrödinger Release 2020-4: Protein Preparation Wizard, Schrödinger, LLC, New York, NY, 2020) was used to optimize the protein for docking. The preparation steps used are standard: addition of hydrogens, optimization of the hydrogen bond networks, and assignment of protonation states of histidine residues. Water molecules were removed, followed by a restrained minimization step using the OPLS4 force field, which uses an implicit water model, generally useful for globular proteins [44]. The Receptor Grid Generation utility of Maestro was then employed to specify the docking site by generating a cubical grid box with the centre at (9.11, 27.14, 22.64) and 23

Å length. This is the centre of the active site, so we are avoiding an examination of allosteric inhibitors or peptides that might be useful in preventing dimerization.

2.2.3 Docking

Glide

In the first step of virtual screening, the docking was performed using the Glide docking tool of Maestro with HTVS (high throughput virtual screening) mode giving one output pose for each ligand [46]. Due to the sheer size of the ligand library, the top 600,000 (~30%) hits were selected for a robust screening step using Glide SP-peptide (standard precision-peptide) docking mode. The top 60,000 (10%) hits were subjected to Glide XP (extra precision) docking.

Gold

The peptide library was also screened using the Gold ChemPLP scoring function. The top 600,000 (~30%) hits were subjected to the Gold fitness score with 10% search efficiency (Goldscore 10%). The top 60,000 (10%) hits were selected for a vigorous screening using a 200% search efficiency of Gold fitness score (Goldscore 200%).

MMGBSA

The top 6000 (10%) hits from Glide XP and the top 6000 (10%) hits from Gold docking were selected for the MMGBSA analysis. The estimated binding free energy of the top hits was calculated with the Prime/MMGBSA (molecular mechanics generalized Born surface area) module using the VSGB solvation model and the OPLS4 force field [46].

Molecular dynamics simulation

The Maestro System Builder utility was employed to prepare docked protein-ligand complexes for molecular dynamics (MD) simulation. The simple point-charge (SPC) water model was used with an orthorhombic box extending 10 Å from protein. The placement of ions was excluded within 20 Å of the ligand. Sodium chloride was added to make the concentration 0.15 M, and extra sodium or chloride ions were added to neutralize the system. The OPLS4 force field was used for simulation. Initially, the NPT ensemble was used at 300 K and 1.01325 bar, and the system was relaxed for 100 ps before simulation. The simulation was recorded for 12 ns with a 3 ps recording interval. The complexes presenting reasonable stability for 12 ns were selected for 100 ns simulation with recording interval of 4 ps (Schrödinger Release 2020-4: Desmond Molecular Dynamics System, D. E. Shaw Research, New York, NY, 2020. Maestro-Desmond Interoperability Tools, Schrödinger, New York, NY, 2020).

3. Results

The overall scheme to build the peptide library is displayed in **Figure 2**. All possible combinations of the twenty canonical amino acids provided 400 dipeptides, 8,000 tripeptides, 160,000 tetrapeptides, and 3.2 million pentapeptides according to the relation 20^n , where n is 2, 3, 4, and 5 for tripeptide, tetrapeptide, and pentapeptide, respectively. Di, tri and tetrapeptides were not subjected to filters due to their smaller size and relative ease of synthesis compared to pentapeptides—they can be reasonably made in the solution phase, and the synthetic challenges associated with solid-phase synthesis can be evaded. Based on aqueous solubility, stability, and ease of synthesis, pentapeptides were passed through a set of seven filters (see methods section above), giving 1,169,013 pentapeptides.

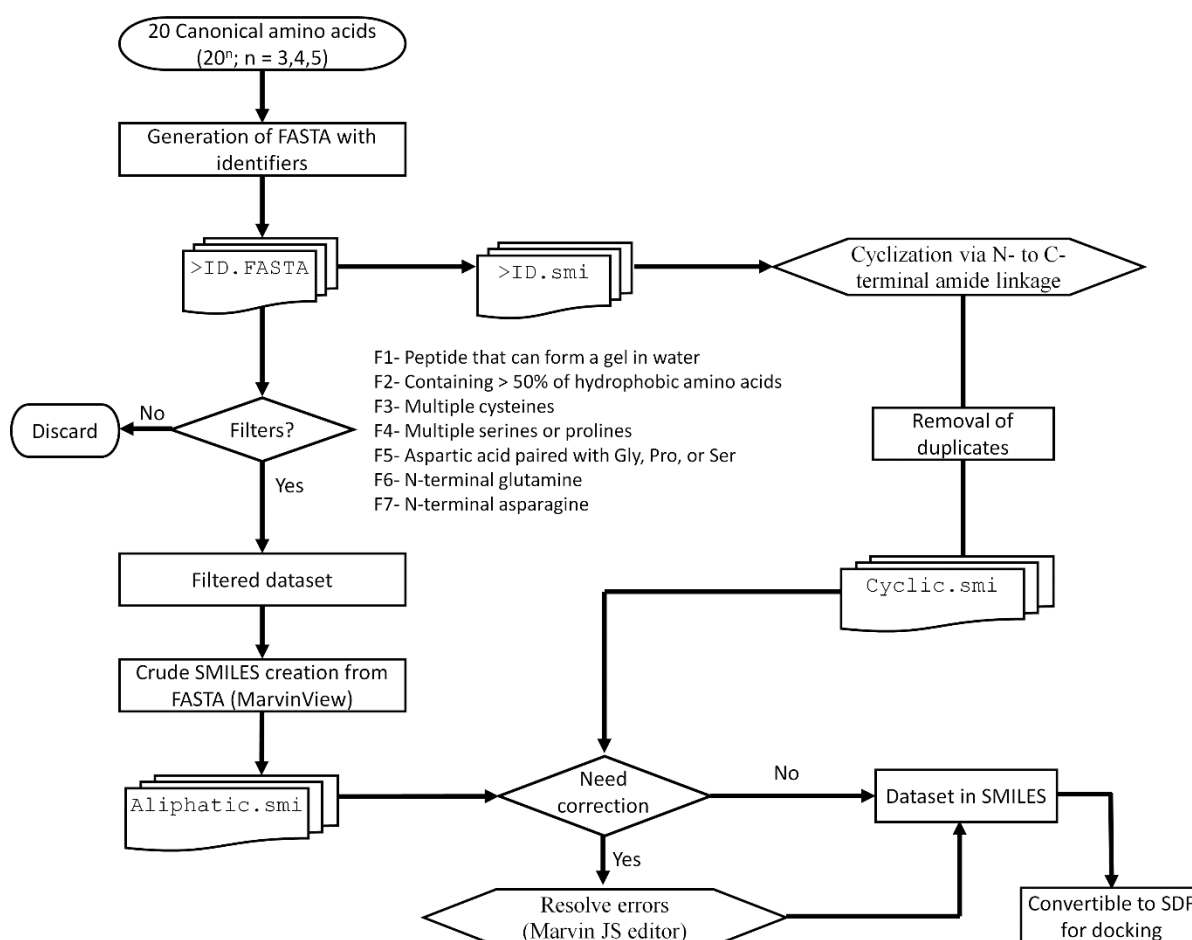


Figure 2: Process diagram displaying the key steps in building the peptide dataset in SMILES and SDF format after applying peptide filters.

Homodetic cyclicpeptides were prepared using 400 dipeptides, 8,000 tripeptides, 160,000 tetrapeptides, and the 3.2 million pentapeptides. N-to-C-terminal amide cyclization of all possible sequences generates substantial number of duplicate peptides. The removal of duplicate entries yielded 210 cyclicdipeptides, 2680 cyclictripeptides, 40,110 cyclictetrapeptides, and 640,016 cyclicpentapeptides.

Collectively, a total of 2,019,819 peptides (aliphatic and cyclic) were generated (**Table 1**). The datasets of aliphatic and cyclicpeptides are available in FASTA and SMILES format in separate files. These files can be readily converted into any desired format (PDB, MOL2, etc.) for molecular docking.

Table 1: Number of peptides generated during dataset construction. The three datasets are available in FASTA, SMILES and SDF formats.

	Dipeptide	Tripeptide	Tetrapeptide	Pentapeptide	Total
Linear	400	8,000	160,000	3,200,000	3,368,400
Cyclic	210	2,680	40,110	640,016	683,016
Linear Filtered	(400)	(8,000)	(160,000)	(1,169,013)	1,337,413
Total	610	10,680	200,110	3,840,016	4,051,416

A total of 2,020,429 peptides (di, tri, tetra and pentapeptides in their linear and cyclic forms) were used for docking against SARS-CoV-2 main protease as a case study to implement the peptide dataset. For computational simplicity, multiple conformations of each peptide were not generated due to the large number of input compounds and the proof-of-principle nature of this work; instead, only one most suitable conformation for each peptide was selected for docking. A more rigorous effort would want to generate 20 low lying conformations per peptide to ensure the conformational space is being sampled.

3.1 Virtual screening of peptide dataset on SARS-CoV-2 main protease

3.1.1 Glide

A quick docking screen of the library using Glide HTVS provided binding energy as low as -10.9 kcal/mol. The top 600,000 hits (~30% of the dataset, -10.9 to -5.1 kcal/mol) were subjected to Glide SP peptide docking, and the top 60,000 hits from this screen (10%, -12.1 to -5.2 kcal/mol) were docked using Glide extra precision (XP) mode. Following this, the top 6000 hits (10%, -13.7 to -8.8 kcal/mol) were subjected to Prime MMGBSA calculations. The MMGBSA data of these “binders” ranged between -90.2 to -9.1 kcal/mol (**Figure 2, 3**).

3.1.2 Gold

An initial screen of the library was performed with the ChemPLP score in Gold. The top 600,000 hits (~30%, 113.5 to 46 ChemPLP score) were docked using Goldscore with 10% efficiency. The top 60,000 hits from this assay (top 10% with 112.6 to 52.2 Goldscore 10% efficiency) were further docked using Goldscore 200% efficiency. The Goldscore 200% values of these hits ranged from 117.3 to 50.1. The top 6000 hits from this step (top 10%, 117.3 to 89.6 Goldscore 200% efficiency) were subjected to Prime MMGBSA. The results from this MMGBSA calculation ranged between -89.3 to -12.4 kcal/mol (**Figure 3, 4**).

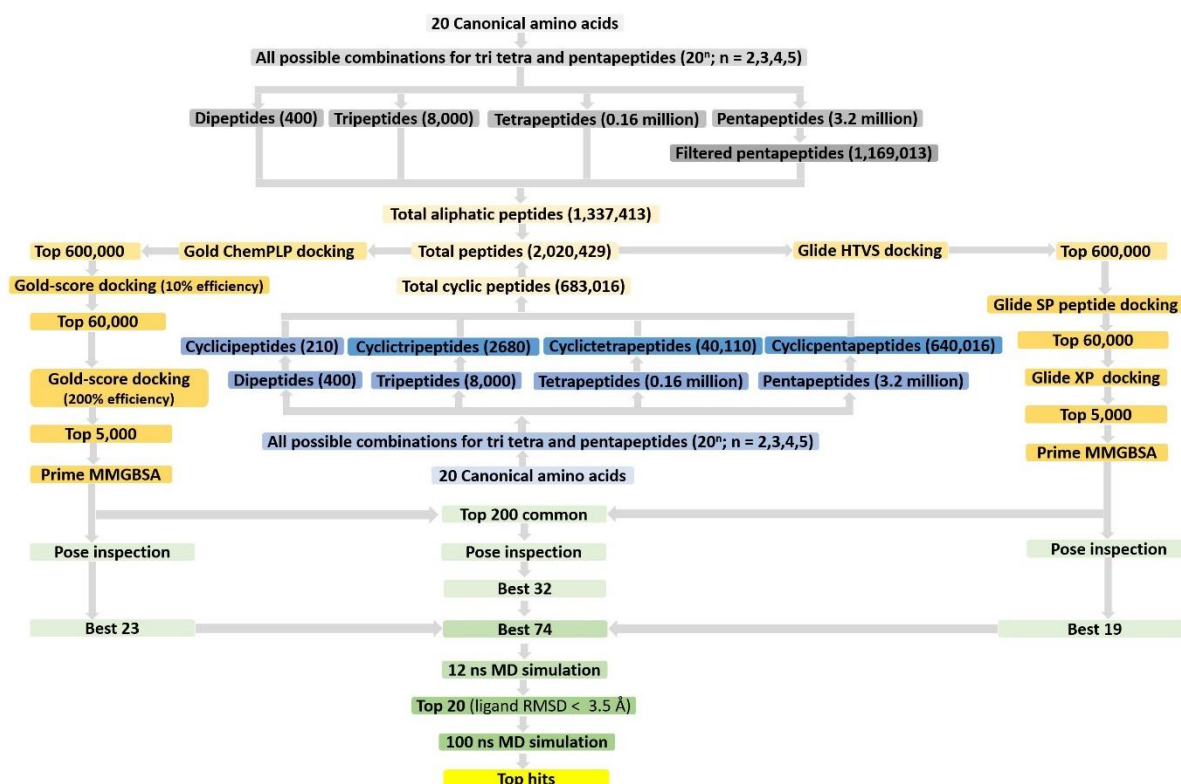


Figure 3: The schematic block diagram of the *in silico* workflow followed in the study to screen peptides against SARS-CoV-2 Main protease. Lighter to darker colour represents the process of a workflow.

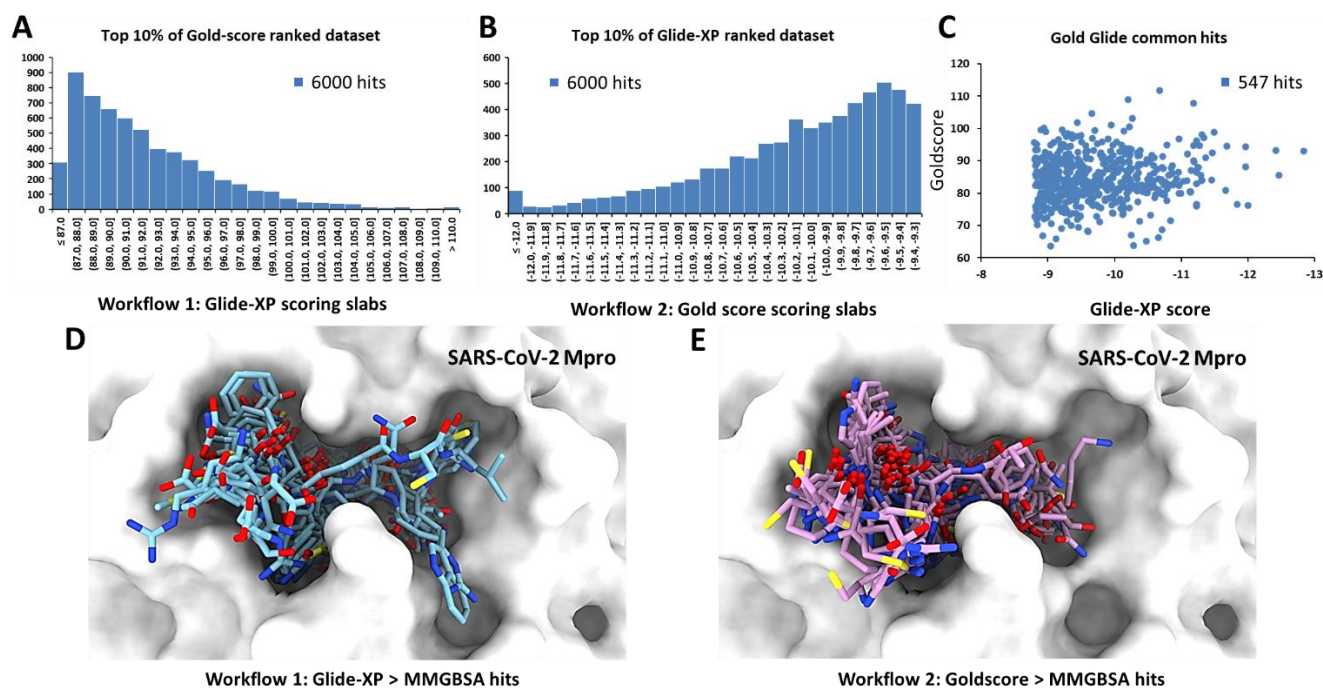


Figure 4: Peptide screening results from two independent workflows. **(A)** Top 6000 (10%) hits obtained from Glide HTVS > SP > XP score, distributed in Glide XP docking score slabs. **(B)** Top 6000 (10%) hits obtained from Gold ChemPLP > Goldscore 10% efficiency > Goldscore 200% efficiency, distributed in Goldscore 200% efficiency slabs. **(C)** Scatter plot of Glide XP docking score vs Goldscore 200% efficiency of the 547 common top hits found from both top 6000 hits from each study taken from corresponding top-ranked datasets. **(D)** Docked conformations of the top hits identified from workflow 1 (Glide HTVS > SP-peptide > XP > Prime-MMGBSA) and **(E)** workflow 2 (Gold ChemPLP score > Goldscore 10% > Goldscore 200% > Prime-MMGBSA) inside the binding pocket of SARS-CoV-2 main protease.

3.1.3 Molecular dynamics simulations

With this integrated parallel virtual screening pipeline, the top hits were selected for MD simulation after careful consideration, completed in 3 steps, 1) top 25 hits from Glide HTVS > Glide SP peptide > Glide XP > Prime MMGBSA workflow; 2) top 25 hits from Gold ChemPLP > Goldscore 10% > Goldscore 200% > Prime MMGBSA workflow, and 3) top 25 common hits that displayed consensus binding poses after both workflows (**Figure 4**). The latter was selected extensively through visual inspection, *i.e.*, peptides that revealed consensus binding conformation by both docking approaches and displayed significant molecular interactions with the critical binding pocket residues, excluding peptides that showed unrealistic docking conformations. These might not be the top hits arising from either algorithm, but a major drawback of single tool-docking studies is their poor correlation to experimental data [19, 28, 47, 48]. Correlation is always improved when molecules are

identified using multiple different workflows, and we consider 2 methods the absolute minimum acceptable. We suggest at least four mutually independent methods for any planned CADD campaign.

These 75 hits were subjected to 12 ns MD simulations to analyse the stability of the complex, interaction energetics, and the conservation of specific intermolecular interactions. After MDs, the following parameters were taken into consideration to select the best hits, 1) a favorable binding affinity (kcal/mol) in terms of MMGBSA calculations; 2) minimum fluctuations of peptide inside the binding pocket (RMSD <3.5 Å); 3) favorable interactions as determined from per-residue decomposition analysis, and 4) having at least one stable H-bond (highest occupancy) with one of the residues of the catalytic dyad over the 12 ns simulation.

The initial poses of all the selected 75 peptides fitted pretty well in the Mpro binding pocket, predominantly interacting with the catalytic dyad. This is to be expected: only the best-fitting peptides would occupy the active site, and a strong hydrogen bond with a member of the catalytic dyad that is already primed to be nucleophilic and acidic was almost going to be a certainty. As the MD simulation progressed, some of the peptides migrated significantly and decreased their interaction with the catalytic dyad. This resulted in an unstable binding conformation; therefore, they could be excluded from further studies. The least disrupted 20 peptides were selected for further analysis and were subjected to a moderate 100 ns MD simulation. Of these, ten protein-peptide complexes displayed a stable RMSD over the simulation period, with the peptides staying in the binding pocket and interacting tightly with the protein's residues. All these peptides displayed significant interactions, including H-bond with His41 and Cys145 of the catalytic dyad (occupancy, >30%; average distance, <3 Å). We discuss the two best-binding peptides (HLFNT and GQYWH); the remainder (GHYFH, FKQH, IKRDM, CHQLN, GQRFL, AQRFS, AQQFN, and EEYCM) is discussed in the supplementary information.

3.1.3.1 HLFNT and GQYWH as high affinity binders for Mpro

An analyses of HLFNT and GQYWH's interactions with Mpro was conducted on the MD simulation data (**Figures 5, 6, and 7**). The final frame of the 12 ns production run was used as the initial frame for the extended 100 ns simulation. The RMSD of the C α -backbone atoms of Mpro with bound peptide was examined and provided information about the overall stability

of the system. For both of these peptides, the Mpro remained stable in the presence of the peptide and the peptide also rapidly converged. RMSD values.

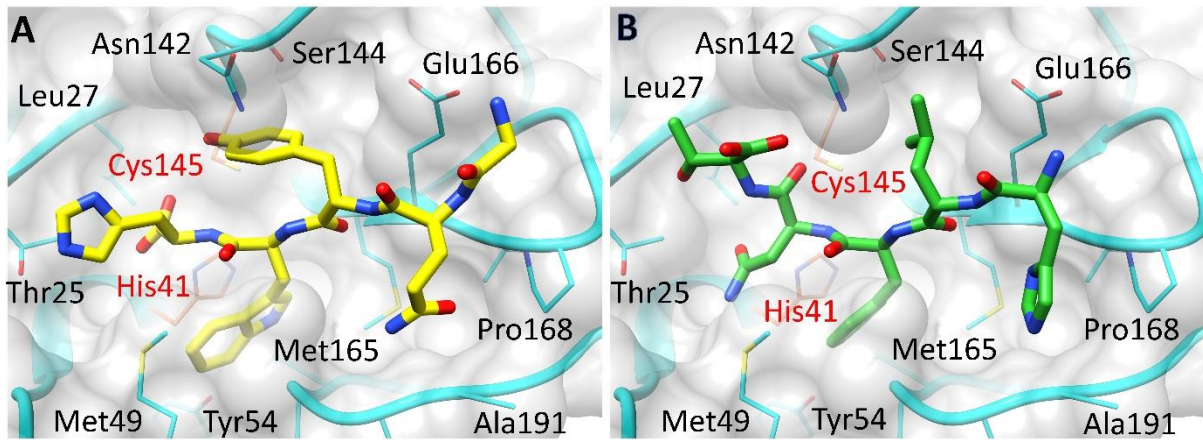


Figure 5: The binding conformation of (A) GQYWH and (B) HLFNT inside the binding pocket of SARS-CoV-2 Mpro. The binding pocket residues are labelled and displayed in stick representation, while catalytic dyad residues are colored red.

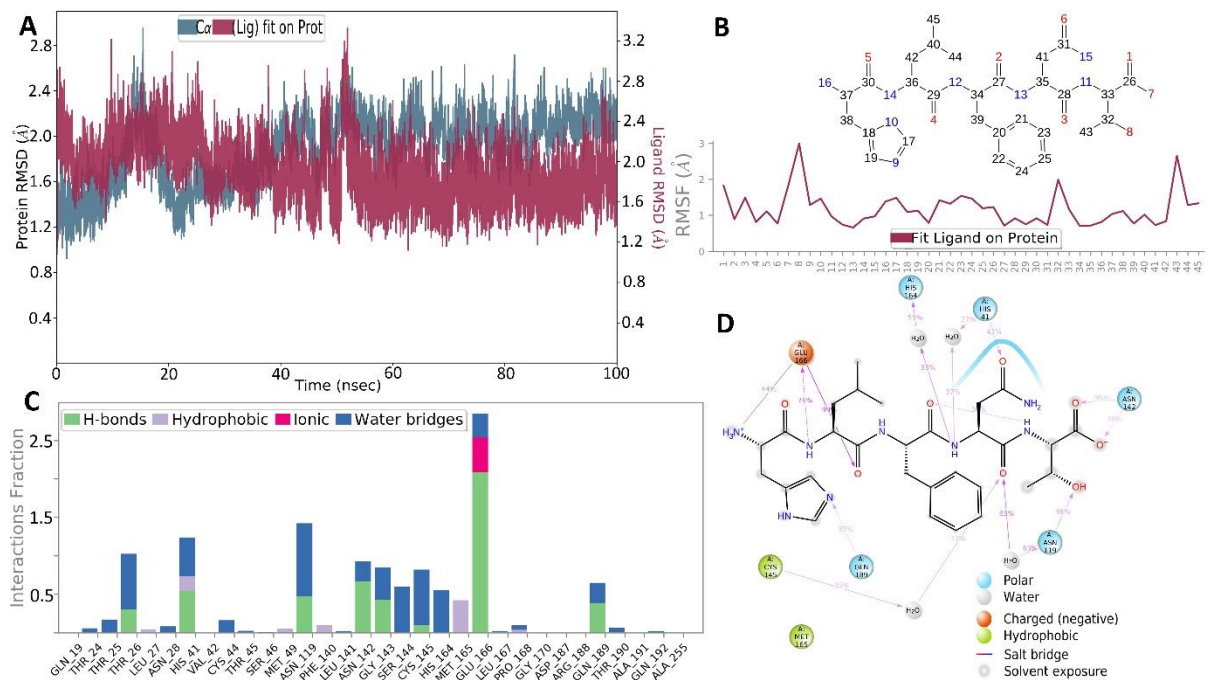


Figure 6: Post-MD simulation analysis of HLFNT@Mpro. (A) The Root Mean Square Deviation (RMSD) of Mpro C α -backbone bound to HLFNT over the simulation period. (B) Ligand-Root Mean Square Fluctuation of HLFNT for characterizing changes in the peptide atom positions during MD simulations. (C) SARS-CoV-2 Mpro interactions with the peptide were monitored throughout the MD simulations. (D) A schematic plot of detailed ligand atom interactions with the protein residues; interactions that occur more than 30% during the simulation period are displayed.

formation over the simulation period was possibly due to the formation of this consistent stacking interaction.

4. Discussion

A dataset of all sequences of linear di, tri, and tetra peptides has been previously published by others to help CADD [49]. This or in-house libraries have been used, and there are reports of efforts to comprehensively screen all possible di or tripeptides against a target protein [50, 51]; however, to the best of our knowledge, no machine-readable database is available with all possible pentapeptides or the N-to-C-terminal cyclic tri, tetra, and pentapeptides. This manuscript is accompanied by the publication of this library for use by others. We also provide a simple method for conducting prefiltration on these peptides. These libraries are provided in FASTA and SMILES formats that can be easily converted to dock-able formats, *e.g.* PDB, MOL2, *etc.* We also provide the SDF file with the prefiltered library we used in this study.

We also chose to prefilter the dataset. It would not take significantly more computational time to screen all possible pentapeptides, but it is important to have a method for arbitrarily reducing the number of simulations to be run according to a research team's whims. The filtrations we conducted here are reasonable to eliminate problematic sequences, but it would also be reasonable for a researcher to not use these filters on a small dataset. The filters were selected in this case so that all peptides arising from the screen would be trivial to synthesize using automated solid phase peptide synthesis [52]. The method of filtration and this filtered data set are provided for ease of use by others should they so desire.

Most small peptides (di, tri, tetra and pentapeptides) have suitable physicochemical properties and are easy to synthesize. The natural zwitterion assists with solubility. However this charge density decreases as the peptide grow in length so some would be sparingly soluble, while other sequences can prove to be unstable.

The chosen filters fall into three categories: solubility, stability, and ease of synthesis. In terms of solubility, we filtered out sequences liable for gel formation (>75% D, E, H, K, N, Q, R, S, T or Y, Filter 1). These would not likely be soluble as distinct species and would likely give false positives on many binding assays. We also removed sequences with a preponderance of aliphatic hydrophobic residues (Filter 2), again the concern is the solubility of these peptides providing false positives in assays due to precipitation and non-specific binding [53]. We then removed a series of notoriously unstable sequences. Cysteine and methionine are liable to rapid oxidation and cleavage of protecting groups during synthesis, making subsequent purification complicated; peptides containing multiple of either sulfur-containing residues were removed (Filter 3). The final category is synthetically challenging sequences that might have reduced purity coming off a bead. A typical pentapeptide needs only a simple precipitation followed by solid phase extraction to render a sufficiently pure

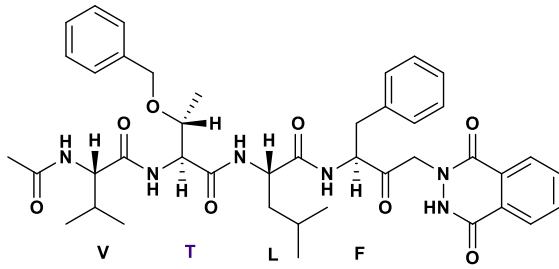
material for evaluation. Some combinations of residues are known to be problematic and could interfere with this workflow. Short peptides containing two or more serines or prolines have a higher probability of leading to deletions during synthesis and were removed (Filter 4). Similarly, aspartic acid residues adjacent to glycine, serine, or proline are susceptible to hydrolysis during acidic cleavage conditions, typically used in solid-phase peptide synthesis (Filter 5). Finally, N-terminal glutamines readily cyclize to pyroglutamate and N-terminal asparagines are irritating to deprotect without decomposition. Peptides with either motif were removed (Filters 6 and 7). As this is a virtual study, these filters could have been omitted, but were included to show the ease of use of the technique and the library manipulation. Even then, we didn't prefilter the tetra-, tri-, or dipeptides in any way (**Figure 3**).

To understand the challenges associated with screening these peptides in a CADD workflow, we selected SARS-CoV-2 Mpro as a potential target due to its compact and flexible catalytic site [54]. Precise binding affinity and peptide conformational prediction of the protein-peptide complex is challenging [47, 55]. We established an integrated peptide virtual screening pipeline to overcome this by incorporating two state-of-the-art molecular docking suits (**Figure 3**). The benchmark studies have reported that the Goldscore is the most efficient for small peptide docking among different scoring functions available in the Gold docking suite [19]. Goldscore can predict near-native conformations of top-scoring peptides (having more than 9 residues) with an accuracy higher than 30% [19]. The Glide SP-peptide has been proven to be 53% accurate in predicting peptide poses. The accuracy could be improved to 58% when SP-peptide is preceded by Prime MMGBSA [56, 57].

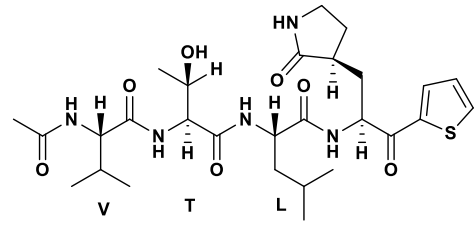
The binding pocket of SARS-CoV-2 Mpro is not deep, and most of it is solvent-exposed. The binding pocket of SARS-CoV-2 Mpro in two apo structures, *7ALH* and *6M2Q*, interact with 12 and 10 water molecules, respectively [58, 59]. The part of the inhibitors sitting in this solvent-accessible region could interact with the solvent, thereby reducing the strength of the protein-ligand interaction. Under these conditions, most docking protocols do not account for the role of water molecules, and docking accuracy can be very poor in highly solvent-accessible binding pockets. The MMGBSA and, especially the MD simulation, becomes very important due to the involvement of water molecules.

Neither of our top hits were previously identified in the literature, nor have they been evaluated against Mpro. Some custom small peptides (**1-8**) analogous to our selected top hits (GQYWH and HLFNT) tested as SARS-CoV Mpro inhibitors are presented in **Figure 8**. The peptides **1**, **2**, **3**, and **4** have been tested to inhibit Mpro at 100 μM with percentage inhibition values of 72%, 23%, <10%, and <10%, respectively [60]. The peptide **5** is a potent inhibitor of SARS-CoV Mpro with K_i of 58 nM [61]. Peptides **6**, **7**, and **8** have been cocrystallized with Mpro of SARS-CoV. The K_i value of **6** was 10.7 μM [62], and the IC_{50} of **7** was 80 μM [63]. The K_{inact}/K_i value of **8** against Mpro of SARS-CoV was 1900 $\text{M}^{-1}\text{s}^{-1}$ [64].

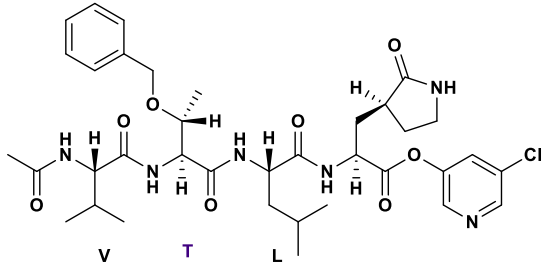
Peptide **8** has a strong structural analogy with our identified hit, HLFNT. The N-terminal histidine of **HLFNT** is replaced with benzyl hydrogen carbonate in **8** and the following two amino acids, leucine and phenylalanine, are identical in both peptides. The fourth asparagine residue of HLFNT is replaced with glutamine type of residue having N in place of α -carbon in **8**, having only one extra methylene residue. The C-terminal threonine of HLFNT is substituted with ethyl ester of 2-hydroxypropanoic acid in **8**, both having hydroxyl groups. This close analogy validates the use of the prepared library and authenticates our *in silico* screening workflow.



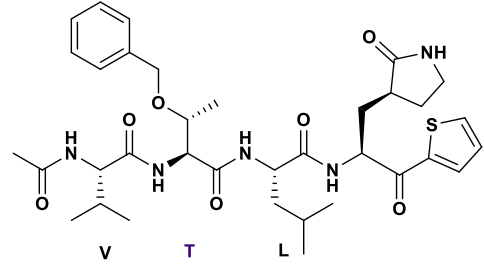
Inhibition = 72% at 100 micromolar
1



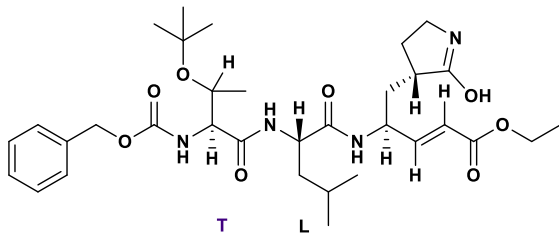
Inhibition = 23% 100 micromolar
2



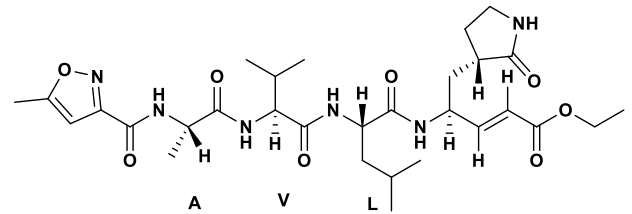
Inhibition < 10% 100 micromolar
3



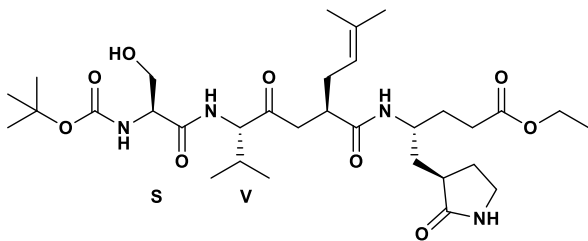
Inhibition < 10% 100 micromolar
4



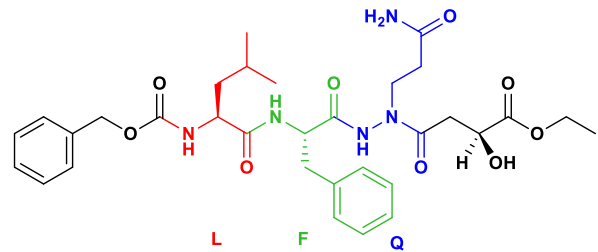
$K_i = 58.0$ nM
5



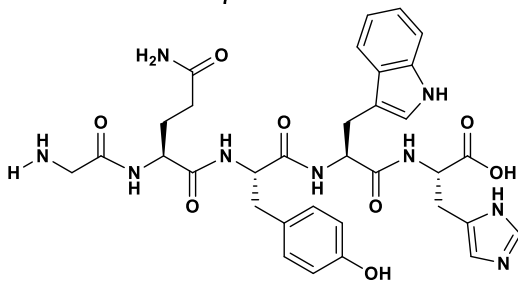
PDB ID of complex 1WOF
PDB ID of ligand: 112
 $K_i = 10.7$ micromolar
6



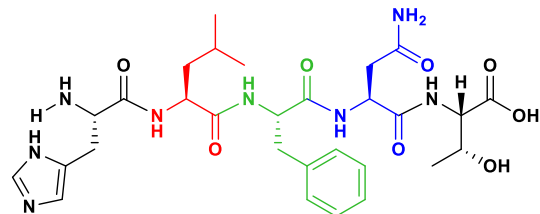
PDB ID of complex: 2QIQ
PDB ID if ligand: CYV
 $IC_{50} = 80$ micromolar
7



PDB ID if complex: 2A5I
PDB ID of ligand: AZP
 $K_{inact}/K_i = 1900$ M⁻¹s⁻¹
8



GQYWH



HLFNT

Figure 8: Structures of small peptidomimetics reported as SARS-CoV Mpro inhibitors. Structures of two top hits (GQYWH and HLFNT) identified in this study. The hydroxyl group of threonine (T) in entries 1, 3, 4, and 5 is in the ether form with benzyl or *tert*-butyl groups.

Conclusion

Among the marketed therapeutics used today, peptides have proved a valuable niche in the drug development spectrum substituting small molecules and large biological therapeutics. The study provides a library of small aliphatic peptides and their N-to-C-terminal cyclized analogues using all possible sequences of 20 canonical amino acids. We have also devised a comprehensive strategy to handle such large peptide datasets utilizing an appropriate array of tools to extract desirable outcomes. The given library could be screened against any desired target protein to identify small peptide hits. These hits could go through the peptidomimetic approaches, *e.g.*, attachment of amino acids on the C or N-terminal to extend the length, use of noncanonical amino acids, *etc.*, for improved activity. This work would profoundly impact the small peptide-based drug discovery domain.

Acknowledgement

The authors highly acknowledge the Natural Sciences and Engineering Research Council of Canada (grant no: 2018-06338) for providing the resources to conduct this work.

Conflict of interest

The authors declare no conflict of interest.

Author contribution

Experimental work and data collection: Sarfraz Ahmad, Muhammad Usman Mirza; design of the study: Sarfraz Ahmad, John F. Trant; manuscript writing: Muhammad Usman Mirza, Sarfraz Ahmad; critical revision of the manuscript: John F. Trant.

Supplementary data

The file “Sample pentapeptides 1st 1 million filtration (all 7 filters).xlsx” contains a sample of the filtration process applied to the 1st 1 million pentapeptides.

The file “Supplementary information.docx” contains MD simulation of top 8 hits with SARS-CoV-2 MPro other than the best two hits, already discussed in the main manuscript.

FASTA, SMILES, and SDF-3D files of all di, tri, tetra, and pentapeptides and their cyclic analogues could be downloaded from the link <https://doi.org/10.5683/SP3/XNRBRN>.

References

1. Bruno, B.J., G.D. Miller, and C.S. Lim, *Basics and recent advances in peptide and protein drug delivery*. Therapeutic delivery, 2013. **4**(11): p. 1443-1467.
2. Anand, U., et al., *Translational aspect in peptide drug discovery and development: An emerging therapeutic candidate*. BioFactors, 2022.
3. Muttenthaler, M., et al., *Trends in peptide drug discovery*. Nature Reviews Drug Discovery, 2021. **20**(4): p. 309-325.
4. Petsalaki, E. and R.B. Russell, *Peptide-mediated interactions in biological systems: new discoveries and applications*. Current opinion in biotechnology, 2008. **19**(4): p. 344-350.
5. Wang, L., et al., *Therapeutic peptides: current applications and future directions*. Signal Transduction and Targeted Therapy, 2022. **7**(1): p. 1-27.
6. Johnston, S.L., *Biologic therapies: what and when?* Journal of clinical pathology, 2007. **60**(1): p. 8-17.
7. Boehncke, W.-H. and N.C. Brembilla, *Immunogenicity of biologic therapies: causes and consequences*. Expert review of clinical immunology, 2018. **14**(6): p. 513-523.
8. Makurvet, F.D., *Biologics vs. small molecules: Drug costs and patient access*. Medicine in Drug Discovery, 2021. **9**: p. 100075.
9. Škalko-Basnet, N., *Biologics: the role of delivery systems in improved therapy*. Biologics: targets & therapy, 2014. **8**: p. 107.
10. Smith, M.C. and J.E. Gestwicki, *Features of protein–protein interactions that translate into potent inhibitors: topology, surface area and affinity*. Expert reviews in molecular medicine, 2012. **14**.
11. de Lomana, M.G., et al., *Consideration of predicted small-molecule metabolites in computational toxicology*. Digital Discovery, 2022. **1**(2): p. 158-172.
12. Waldmann, H., *Human monoclonal antibodies: the residual challenge of antibody immunogenicity*. Human Monoclonal Antibodies, 2014: p. 1-8.
13. Fosgerau, K. and T. Hoffmann, *Peptide therapeutics: current status and future directions*. Drug discovery today, 2015. **20**(1): p. 122-128.
14. Hale, M., et al., *Basic tetrapeptides as potent intracellular inhibitors of type A botulinum neurotoxin protease activity*. Journal of Biological Chemistry, 2011. **286**(3): p. 1802-1811.
15. Vlieghe, P., et al., *Synthetic therapeutic peptides: science and market*. Drug discovery today, 2010. **15**(1-2): p. 40-56.
16. Craik, D.J., et al., *The future of peptide-based drugs*. Chemical biology & drug design, 2013. **81**(1): p. 136-147.
17. Kaspar, A.A. and J.M. Reichert, *Future directions for peptide therapeutics development*. Drug discovery today, 2013. **18**(17-18): p. 807-817.
18. Weng, G., et al., *Comprehensive evaluation of fourteen docking programs on protein–peptide complexes*. Journal of chemical theory and computation, 2020. **16**(6): p. 3959-3969.
19. Hauser, A.S. and B.r. Windshügel, *LEADS-PEP: a benchmark data set for assessment of peptide docking performance*. Journal of chemical information and modeling, 2016. **56**(1): p. 188-200.
20. Steuten, K., et al., *Challenges for targeting SARS-CoV-2 proteases as a therapeutic strategy for COVID-19*. ACS infectious diseases, 2021. **7**(6): p. 1457-1468.
21. Anand, K., et al., *Coronavirus main proteinase (3CLpro) structure: basis for design of anti-SARS drugs*. Science, 2003. **300**(5626): p. 1763-1767.
22. Zhao, Y., et al., *Crystal structure of SARS-CoV-2 main protease in complex with protease inhibitor PF-07321332*. Protein & Cell, 2022. **13**(9): p. 689-693.
23. Lee, J., et al., *Crystallographic structure of wild-type SARS-CoV-2 main protease acyl-enzyme intermediate with physiological C-terminal autoprocessing site*. Nature Communications, 2020. **11**(1): p. 5877.
24. Zhang, L., et al., *Crystal structure of SARS-CoV-2 main protease provides a basis for design of improved α -ketoamide inhibitors*. Science, 2020. **368**(6489): p. 409-412.

25. Zhang, L., et al., *α -Ketoamides as broad-spectrum inhibitors of coronavirus and enterovirus replication: structure-based design, synthesis, and activity assessment*. Journal of medicinal chemistry, 2020. **63**(9): p. 4562-4578.
26. Ahmad, S., et al., *Fragment-based in silico design of SARS CoV-2 main protease inhibitors*. Chemical Biology & Drug Design, 2021.
27. Wu, C., et al., *Analysis of therapeutic targets for SARS-CoV-2 and discovery of potential drugs by computational methods*. Acta Pharmaceutica Sinica B, 2020. **10**(5): p. 766-788.
28. 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*. Medicinal Research Reviews, 2022. **42**(2): p. 744-769.
29. Breidenbach, J., et al., *Targeting the Main Protease of SARS-CoV-2: From the Establishment of High Throughput Screening to the Design of Tailored Inhibitors*. Angewandte Chemie International Edition, 2021. **60**(18): p. 10423-10429.
30. Pillaiyar, T., et al., *Small-Molecule Thioesters as SARS-CoV-2 Main Protease Inhibitors: Enzyme Inhibition, Structure–Activity Relationships, Antiviral Activity, and X-ray Structure Determination*. Journal of Medicinal Chemistry, 2022. **65**(13): p. 9376-9395.
31. Hu, Q., et al., *The SARS-CoV-2 main protease (Mpro): Structure, function, and emerging therapies for COVID-19*. MedComm, 2022. **3**(3): p. e151.
32. La Monica, G., 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. **65**(19): p. 12500-12534.
33. Gao, K., et al., *Perspectives on SARS-CoV-2 Main Protease Inhibitors*. Journal of Medicinal Chemistry, 2021. **64**(23): p. 16922-16955.
34. Bzówka, M., et al., *Structural and Evolutionary Analysis Indicate That the SARS-CoV-2 Mpro Is a Challenging Target for Small-Molecule Inhibitor Design*. International Journal of Molecular Sciences, 2020. **21**(9): p. 3099.
35. Quan, B.-X., et al., *An orally available Mpro inhibitor is effective against wild-type SARS-CoV-2 and variants including Omicron*. Nature Microbiology, 2022. **7**(5): p. 716-725.
36. Kashyap, P., et al., *A ricin-based peptide BRIP from Hordeum vulgare inhibits Mpro of SARS-CoV-2*. Scientific Reports, 2022. **12**(1): p. 12802.
37. Johansen-Leete, J., et al., *Antiviral cyclic peptides targeting the main protease of SARS-CoV-2*. Chemical Science, 2022. **13**(13): p. 3826-3836.
38. *Marvin was used for the conversion of formats of the chemical structures, Marvin version 22.18.0, ChemAxon (<https://www.chemaxon.com>)*.
39. Anderson, E., G.D. Veith, and D. Weininger, *SMILES, a line notation and computerized interpreter for chemical structures*. 1987: US Environmental Protection Agency, Environmental Research Laboratory.
40. Lipman, D.J. and W.R. Pearson, *Rapid and sensitive protein similarity searches*. Science, 1985. **227**(4693): p. 1435-1441.
41. O'Boyle, N.M., et al., *Open Babel: An open chemical toolbox*. Journal of cheminformatics, 2011. **3**(1): p. 1-14.
42. The Open Babel Package, v.h.o.o.a.O.
43. Sastry, G.M., et al., *Protein and ligand preparation: parameters, protocols, and influence on virtual screening enrichments*. Journal of Computer-Aided Molecular Design, 2013. **27**(3): p. 221-234.
44. Roos, K., et al., *OPLS3e: Extending force field coverage for drug-like small molecules*. Journal of Chemical Theory and Computation, 2019. **15**(3): p. 1863-1874.
45. Anson, B., A.K. Ghosh, and A. Mesecar, *X-ray Structure of SARS-CoV-2 main protease bound to GRL-024-20 at 1.45 Å, 6XR3*, in Protein Databank. 2020.

46. Li, J., et al., *The VSGB 2.0 model: a next generation energy model for high resolution protein structure modeling*. Proteins: Structure, Function, and Bioinformatics, 2011. **79**(10): p. 2794-2812.
47. Agrawal, P., et al., *Benchmarking of different molecular docking methods for protein-peptide docking*. BMC bioinformatics, 2019. **19**(13): p. 105-124.
48. Sanner, M.F., et al., *Improving Docking Power for Short Peptides Using Random Forest*. Journal of Chemical Information and Modeling, 2021. **61**(6): p. 3074-3090.
49. Prasasty, V.D. and E.P. Istyastono, *Data of small peptides in SMILES and three-dimensional formats for virtual screening campaigns*. Data in brief, 2019. **27**: p. 104607.
50. Panyayai, T., et al., *The potential peptides against angiotensin-I converting enzyme through a virtual tripeptide-constructing library*. Computational biology and chemistry, 2018. **77**: p. 207-213.
51. Mollica, A., et al., *Combinatorial peptide library screening for discovery of diverse α -glucosidase inhibitors using molecular dynamics simulations and binary QSAR models*. Journal of Biomolecular Structure and Dynamics, 2019. **37**(3): p. 726-740.
52. Petrou, C. and Y. Sarigiannis, *Peptide synthesis: Methods, trends, and challenges*. Peptide applications in biomedicine, biotechnology and bioengineering, 2018: p. 1-21.
53. Sarma, R., et al., *Peptide solubility limits: backbone and side-chain interactions*. The Journal of Physical Chemistry B, 2018. **122**(13): p. 3528-3539.
54. Rocha, R.E., et al., *A higher flexibility at the SARS-CoV-2 main protease active site compared to SARS-CoV and its potentialities for new inhibitor virtual screening targeting multi-conformers*. Journal of Biomolecular Structure and Dynamics, 2021: p. 1-21.
55. Ciemny, M., et al., *Protein-peptide docking: opportunities and challenges*. Drug discovery today, 2018. **23**(8): p. 1530-1537.
56. Tubert-Brohman, I., et al., *Improved docking of polypeptides with Glide*. Journal of chemical information and modeling, 2013. **53**(7): p. 1689-1699.
57. Feher, M. and C.I. Williams, *Numerical errors and chaotic behavior in docking simulations*. Journal of chemical information and modeling, 2012. **52**(3): p. 724-738.
58. Su, H.-x., et al., *Anti-SARS-CoV-2 activities in vitro of Shuanghuanglian preparations and bioactive ingredients*. Acta Pharmacologica Sinica, 2020. **41**(9): p. 1167-1177.
59. <https://www.rcsb.org/structure/7alh>.
60. Zhang, J., et al., *Design, synthesis, and evaluation of inhibitors for severe acute respiratory syndrome 3C-like protease based on phthalhydrazide ketones or heteroaromatic esters*. Journal of medicinal chemistry, 2007. **50**(8): p. 1850-1864.
61. Yang, S., et al., *Synthesis, crystal structure, structure- activity relationships, and antiviral activity of a potent SARS coronavirus 3CL protease inhibitor*. Journal of medicinal chemistry, 2006. **49**(16): p. 4971-4980.
62. Yang, H., et al., *Design of wide-spectrum inhibitors targeting coronavirus main proteases*. PLoS Biol, 2005. **3**(10): p. e324.
63. Ghosh, A.K., et al., *Structure-based design, synthesis, and biological evaluation of peptidomimetic SARS-CoV 3CLpro inhibitors*. Bioorganic & medicinal chemistry letters, 2007. **17**(21): p. 5876-5880.
64. Lee, T.-W., et al., *Crystal structures of the main peptidase from the SARS coronavirus inhibited by a substrate-like aza-peptide epoxide*. Journal of molecular biology, 2005. **353**(5): p. 1137-1151.