The Third CACHE Challenge – Finding Ligands Targeting the Macrodomain of SARS-CoV-2 NSP3 Using AI-inspired and Knowledge-Based Approaches.

Suneel Kumar B.V.S,[1,] Ophélie Rostaing,[1] Mihai Burai-Patrascu,[1] Matej Janezic,[1] Antoine Moitessier,[1] Joshua Pottel,[1] Nicolas Moitessier[1]*

[1] Molecular Forecaster Inc., 910-2075 Robert Bourassa St., Montreal, Quebec, H3A2L1, Canada.

* Corresponding Author: Suneel Kumar B.V.S, suneel.kumar@molecularforecaster.com

Abstract

The SARS-CoV-2 virus contains a host of nonstructural proteins (NSPs) that contribute to its structure and viral function. Among them is the nonstructural protein 3 (NSP3), which contains a macrodomain (Mac1) that interferes with antiviral adenosine diphosphate (ADP)-ribosylation signaling. Catalytic mutations in Mac1 render viruses nonpathogenic, making this enzyme a promising target for antiviral development. For this reason, the third CACHE challenge focused on identifying binders of the Mac1 domain of NSP3 for the development of novel antivirals against SARS-CoV-2. To this end, we used available structural data of the NSP3 Mac1 domain in complex with known fragment binders as starting points for ligand discovery; our efforts were primarily focused on sub-sites of the ADP binding site in the NSP3 macrodomain. Then, using Artificial intelligence (AI)-guided and knowledge-based fragment merging and expansion approaches, we generated novel molecules that would serve as templates to identify highly similar compounds in the Enamine REAL database that would be commercially available**.** Our design yielded a library of 12,800 molecules, which was docked with our program FITTED to a representative crystal structure of NSP3. We ranked the predicted binding poses based on docking score, followed by visual pose analysis of the best 200 compounds. We finally selected and proposed 150 compounds for testing, followed by further shortlisting to yield a final list of 107 molecules. 91 compounds were purchased from Enamine and are being tested at the Structural Genomics Consortium (SGC). Our approach and findings will further contribute to our open science efforts, and we aim to continue to engage the scientific community.

1. Introduction

Coronaviruses, including SARS-CoV-2, are a diverse family of positive-sense single-stranded RNA (+ssRNA) viruses that are capable of infecting mammals, among other species.[1] In the host system, the +ssRNA is translated into two polyproteins – pp1a and pp1ab – which are further post-processed into nonstructural proteins (NSPs) by the papain-like protease (PL^{pro}) and the 3C-like protease ($3CL^{pro}$). These NSPs are essential for the structure of the viral replication and transcription complex.[2] Most efforts to develop SARS-CoV-2 antivirals have focused primarily on targeting PL^{pro} [3, 4] and $3CL^{pro}$ [5–8] (Figure 1), which culminated in the marketing of nirmatrelvir (active ingredient of Paxlovid™, Pfizer) and ensitrelvir (Xocora®, Shionogi) as orally-available medications inhibiting $3CL^{pro}$. [9,10]

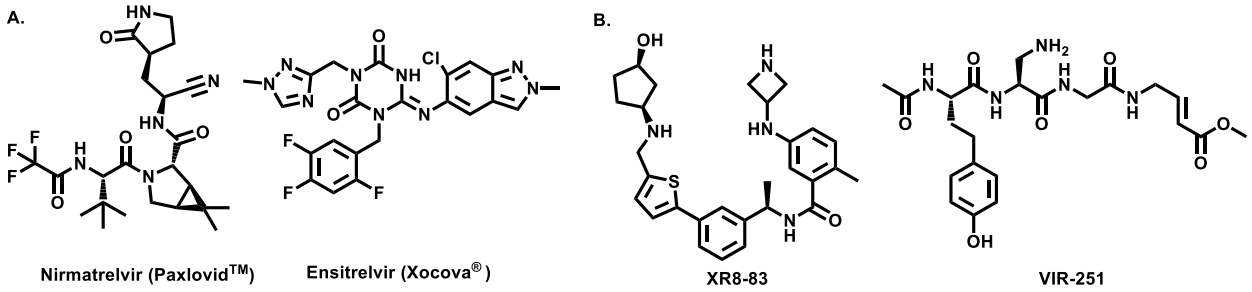

Figure 1. Selected **A)** $3CL^{pro}$ and **B)** PL^{pro} inhibitors. [11–14]

However, antiviral resistance is one major concern related to targeting $3CL^{pro}$, afforded by viral mutations. Indeed, several $3CL^{pro}$ mutations have been associated with a significant decrease in Paxlovid and Xocora inhibitory activity.[13,14] Thus, an alternative approach is to focus efforts on other key proteins. One promising candidate is NSP3, which contains a highly conserved macrodomain (Mac1) that interferes with adenosine diphosphate(ADP)-ribosylation signaling, a key component in mediating the human immune response.[15] This domain contains several binding sites, including the catalytic site and the ADP sub-site. Due to the high degree of Mac1 conservation across all species,[16] it has been posited that catalytic mutations in the Mac1 domain would lead to loss of viral pathogenicity.[17,18] This approach should then circumvent the resistance issues faced by $3CL^{pro}$ and PL^{pro} inhibitors.

Recent work to identify fragment binders of the NSP3 Mac1 domain has led to an explosion of structural data, with over 200 crystal structures deposited in the Protein Data Bank (PDB).[16,19] These comprehensive crystallographic data have marked a turning point in the development of Mac1 inhibitors, leading to the identification of several (low) micromolar binders (Figure 2).

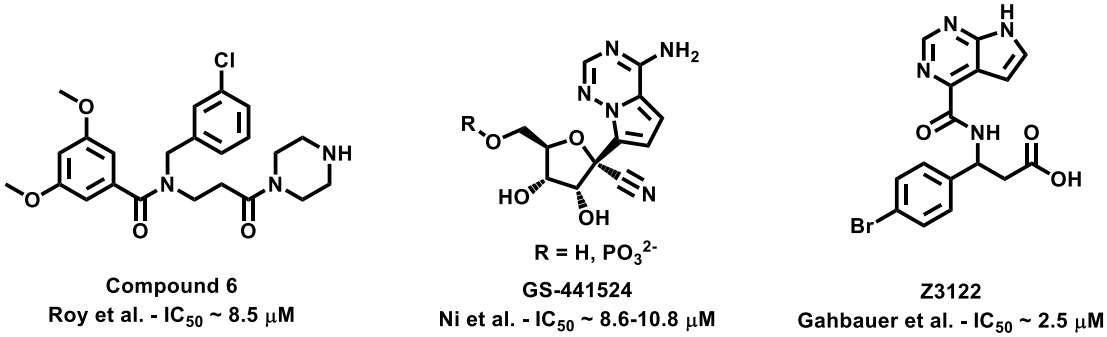

Figure 2. Selected SARS-CoV-2 NSP3 Mac1 domain binders.

Despite continued research, there remains a limited number of antiviral drugs on the market that can serve as alternatives to current treatments for SARS-CoV2. Drugs such as Remdesivir, Molnupiravir, and Paxlovid have shown

varying degrees of effectiveness, ranging from moderate to less effective and some of the reported compounds still display a degree of toxicity .[20–22] Thus, it is crucial to develop novel chemical series that not only demonstrate increased potency but also address previously identified issues of toxicity and resistance. With this goal in mind, we aimed to participate in the CACHE challenge #3, where we could leverage our medicinal and computational chemistry expertise along with the available molecular and structural data to identify novel Mac1 binders with enhanced potency.

2. Approach

Our roadmap begins with the analysis of the structural characteristics of the Mac1 domain. There are over 200 Mac1 crystal structures complexed with small molecules and fragments that are available in the PDB, provides ample information for identifying and developing new Mac1 binders. We thoroughly analyzed the co-crystal fragments and their binding mode as well as Mac1 domain binding sites, protein and ligand flexibility, and key interactions patterns using our in-house Python scripts. For fragment-based merging and expansion studies, we have selected diverse co-crystal fragments binding to the ADP sub-site to design potential Mac1 binders (Figure 3).

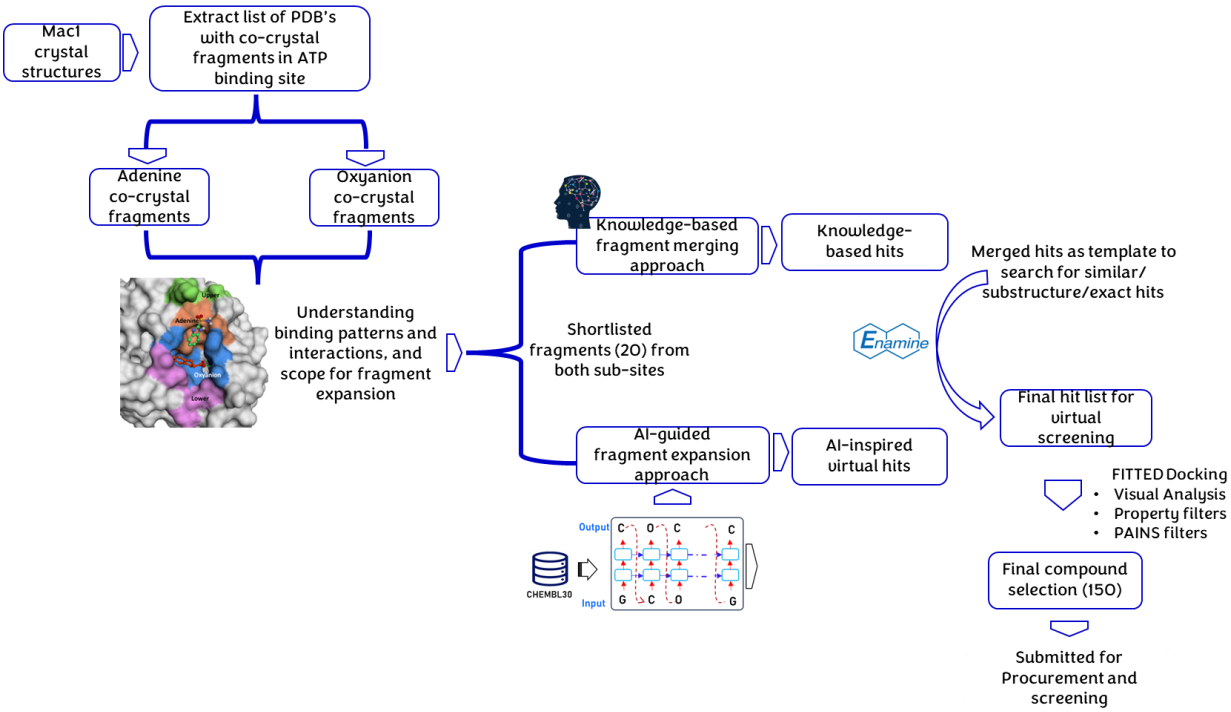

Figure 3. Roadmap for designing novel Mac1 binders using fragment-based approaches.

With the fragment-based expansion approach, we implemented a recurrent neural network (RNN-LSTM)-based generative model which is pre-trained on the CHEMBL database[23–25] before fine-tuning it with known Mac1 binders. A pre-trained model leverages its learned patterns, weights, and knowledge to predict and construct new molecular structures irrespective of specific targets. A fine-tuned model was used for fragment expansion studies to generate molecules from the selected fragments and their corresponding positions. This process is repeated until all suitable positions for the selected fragments are addressed and potential molecules generated using the fine-tuned model. Furthermore, we have leveraged the chemical intuition and expertise of our team to manually select fragments for merging to generate ideas in addition to the generative AI approach. Once the ideation process was completed from

both approaches, we considered generated molecules as templates, to source Enamine REAL database to source analogues, either by substructure or similarity searches.

Following the proposed roadmap, we processed the retrieved the list of compounds ready for docking taking advantage of the various programs integrated into our drug discovery platform FORECASTER.[28] We virtually screen this set of small molecules against a representative crystal structure of the Mac1 domain with our docking program FITTED. The top 200 compounds ranked by docking score were visualized. We focused on 1) overall molecule fit in the ADP sub-site, 2) interaction pattern between protein and ligand consistent with co-crystallized fragments, and 3) structural diversity of ligands. Finally, as per the CACHE3 guidelines, we have shortlisted 150 compounds for purchase and testing.

3. Results and Discussion

A total of 455 crystal structures for Mac1 have been reported and deposited in the protein data bank (PDB), which is being investigated by a variety of research teams (as of February 2023).[16,19,26–28] Of 455 structures, 176 of Mac1 structures were complexed with small molecules, 249 with fragments, and the remainder were reported as either unbound or apo structures. For our study, we specifically considered those Mac1 crystal structures which are complexed with fragments, since these structures are particularly relevant to our research.

3.1. Structural analysis of the active site

A majority of the fragments co-crystallized with Mac1 were bound to the adenosine diphosphate ribose (ADPr) binding region while some were binding allosteric cavities. Dr. Fraser and colleagues have identified and termed specific regions as the Adenine, Oxyanion, Upper, and Lower sub-sites[16] as highlighted in Figure 4. We used in-house python scripts to extract and filter the interactions observed between the co-crystal fragments and the respective Mac1 sub-sites. Following this interaction analysis with four known sub-sites, we refined our selection to 228 co-crystal structures.

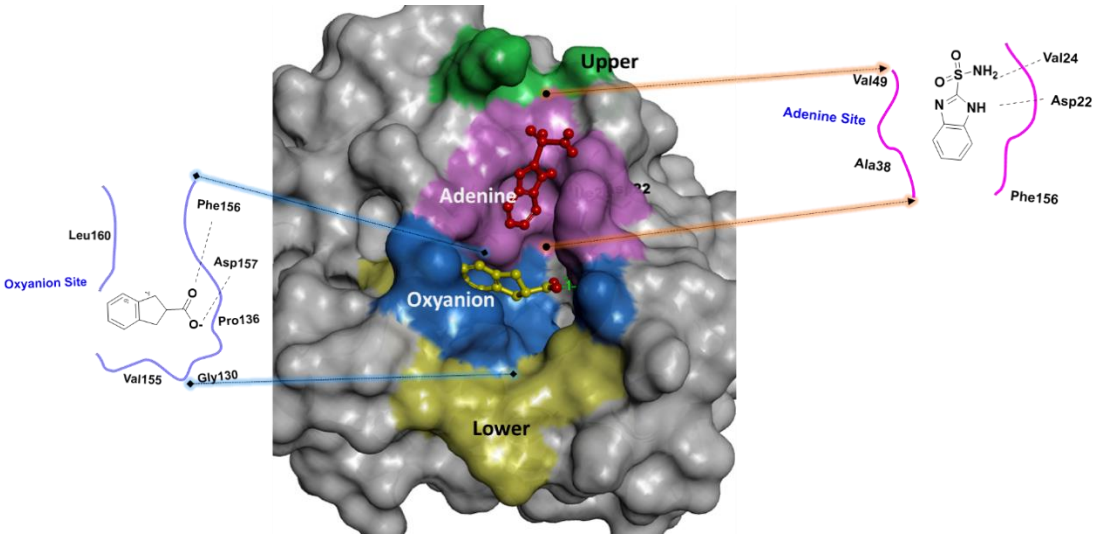

Figure 4: SARS-CoV-2 NSP3 macrodomain (Mac1) crystal structures (PDB: 5RTC, and 5RSW), protein structures are displayed in surface mode and co-crystal fragments were highlighted in ball and stick mode. Various sub-sites on the macrodomain surface are distinctly color-marked: the Adenine sub-site is in pink, the Oxyanion sub-site in blue, the Lower sub-site in yellow, and the Upper sub-site in green. Additionally, 2D-interaction diagram showcases adenine (PDB ID: 5RTC) and oxyanion (PDB ID: 5RSW) co-crystal fragments interacting with the Mac1 domain and highlights the crucial amino acids participating in these interactions.

The majority of these co-crystal fragments (122) bind to the adenine sub-site; they are involved in key hydrogen bonding interactions with Asp22 and Ile23 and they form stacking interactions with Phe156. Additionally, 62 co-crystal fragments were found in the oxyanion sub-site, which is shaped by the backbone NH groups of Phe156 and Asp157. Our study focused on these two sub-sites because of their pocket size, the presence of key interactions, the potential scope for fragment expansion and generating hit-like molecules, and the presence of key residues, as shown in Figure 4.

3.1.1 Adenine sub-site. Of the 228 structures, 122 Mac1 structures contained fragments bound to the adenine sub-site which we termed as adenine co-crystal fragments from here on. Most of the amino acids in the adenine sub-site are hydrophobic, including Phe156, Ile23, Val24, Gly48, Val49, and Ala52, as well as the polar Asp22. A majority of adenine co-crystal fragments form hydrogen bonds with Asp22, Ile23, Val49, while Phe156 is also involved in π-stacking and π-cation interactions. Adenine co-crystal fragments typically had three key pharmacophoric features: 1) a hydrogen bond acceptor, 2) a hydrogen bond donor within 3) a heteroaromatic core (e.g., thiazoles, pyridines, pyrazoles, pyrroles). The frequency of the key interactions observed in adenine sub-site and it's cocrystal fragments are shown in Figure SF1A.

3.1.2 Oxyanion sub-site: Oxyanion co-crystal fragments refer to the fragments that are bound to the oxyanion sub-site of the Mac1 crystal structures (62). The oxyanion site is adjacent to the adenosine sub-site. Many of these fragments (e.g., carboxylates, sulfones, isoxazoles, keto acids, and succinimides) contain highly polar functional groups. The oxyanion sub-site is composed primarily of hydrophobic amino acids, as well as exposed NH groups on Phe156 and Asp157's backbones. As oxyanion holes stabilize negatively charged intermediates or transition states of enzymatic reactions, this is somewhat expected. Almost all of the oxyanion co-crystal fragments hydrogen bonded with Phe156 and Asp157, and some directed groups overlap or extend toward the phosphate tunnel, others toward the lower site, and some overlap with the upper adenine sub-site. The frequency of the key interactions observed in oxyanion sub-site and it's cocrystal fragments are shown in Figure SF1B.

3.2 Hit Discovery through a fragment-based approach:

3.2.1 Selection of potential fragments as starting points for inhibitor design:

Identification of a promising starting point for fragment linking or expansion is key for success in molecular design. In our fragment selection process, we carefully considered several factors, including the rule of three[35] (molecular weight of the fragment < 300, LogP ≤ 3, number of hydrogen bond donors (HBDs) ≤ 3, and number of hydrogen bond acceptors (HBAs) ≤ 3), and number of rotatable bonds. These properties were computed using SMART, a module of the Forecaster suite. We carefully evaluated each fragment for their binding pose and interaction patterns with their respective sub-sites, as well as their orientation and scope for fragment-based studies. Leveraging our team's expertise, we carefully selected twenty starting anchor fragments that are suitable for strategies involving fragment expansion and merging. Out of twenty shortlisted fragments, thirteen of these were associated with the adenine sub-site and the remaining seven with the oxyanion sub-site. To conduct the fragment-based design, we employed two distinct approaches. The first strategy was based on an AI-guided fragment expansion, anchored by established active fragment for a more targeted foundation. In contrast, the second method utilized knowledge-driven techniques for fragment merging. These two approaches are detailed below.

3.2.2 AI-guided fragment expansion:

We implemented a deep recurrent neural network (RNN) [29] with long short-term memory (LSTM) cells for generation to design potential hit candidates for Mac1. RNNs have been instrumental in sequence modelling tasks, which predict or generate sequences of data (e.g., words making up sentences) where the temporal order or sequence of the data points is vital. RNN-LSTM models have found exciting applications in the field of generative AI, such as text

generation, music generation, art generation, and also for drug discovery.[30–33] A filtered ChEMBL30 dataset, a repository rich in chemical diversity, was used to train the RNN-LSTM model initially. We trained our RNN-LSTM model on SMILES (Simplified Molecular Input Line Entry System), a textual representation of molecular structures, to accurately capture the chemical data. The dataset provided valuable insight into a wide variety of chemical structures and their properties. During this early stage of model development, it is crucial to learn the probability distribution of SMILES strings, which is crucial to the model's ability to process and generate chemical data accurately.

For pre-training, the RNN-LSTM model was configured with 256 LSTM units and trained over 22 epochs to balance exposure to the dataset against the risk of overfitting. The Adam optimizer was employed for its adaptability across complex patterns. With a batch size of 256 and a 10% validation split, the model's progress was consistently monitored, leveraging verbose training feedback, checkpoints guided by validation loss. This model will be referred to as the pre-trained model.

Following initial training, the models were fine-tuned through transfer learning, which refined their pre-trained capabilities for the specific task of identifying Mac1 inhibitors (200).[16,19,27] This fine-tuning adjusted the models' weights and biases to focus on the distinct properties of the 200 known Mac1 inhibitors (refer to supporting files). At this stage, the model is now trained to generate realistic molecules which have properties similar to known Mac1 inhibitors. A sampling temperature of 0.75 and a maximum SMILES length of 128 characters were set to balance the generation of diverse and complex molecular structures without taxing the model's capacity. This model will be referred to as the fine-tuned model.

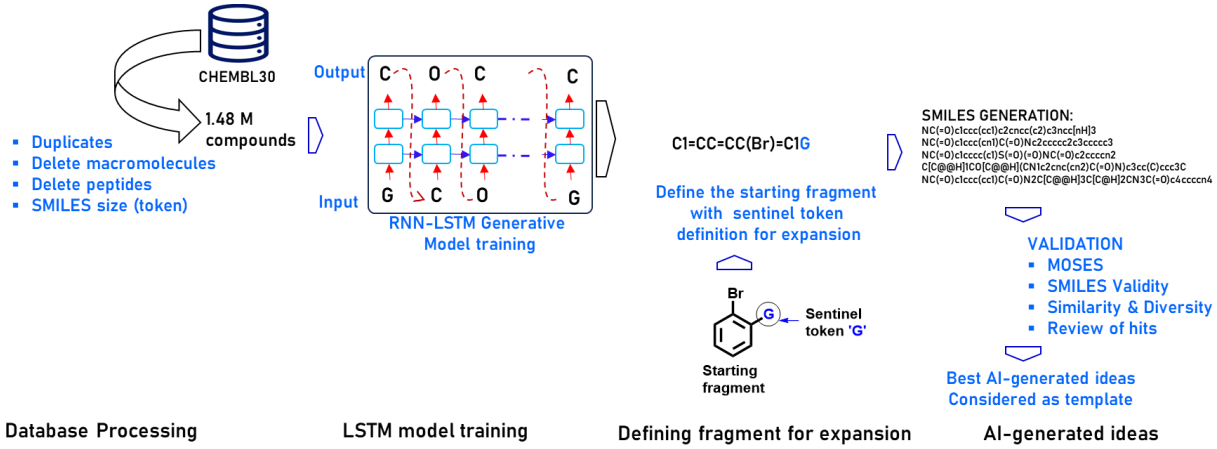

Figure 5: Schematic diagram of RNN-LSTM model training and sampling process. To explain the sampling process, here we considered bromobenzyl as the initial fragment, the 2nd position is marked by the 'G' sentinel token, directing model expansion at this location.

The process of AI-guided molecular design (RNN-LSTM) begins with a sentinel token known as ‘G’. This token will be used in the initial sampling process (molecule generation) followed by a transfer learning approach for fine-tuning the hits in the second stage. Figure 5 illustrates the molecule generation process, which begins with a shortlisted fragment (called a "primer") and uses the sentinel token ‘G’ to indicate where the fragment needs to be expanded or grown. We selected seven of the thirteen adenine co-crystal fragments (PDB IDs: 5RVI, 5RTE, 5RSC, 5RUD, 5RTN, 5RTM, and 5S2G) and two of the seven oxyanion co-crystal fragments (PDB IDs: 5RSP, 5RTG) for our study (Figure 6) as starting points for this approach. The RNN-LSTM utilizes its learned knowledge (from training) to iteratively build upon the fragment, improving compound design within the desired chemical space of Mac1 inhibitors.

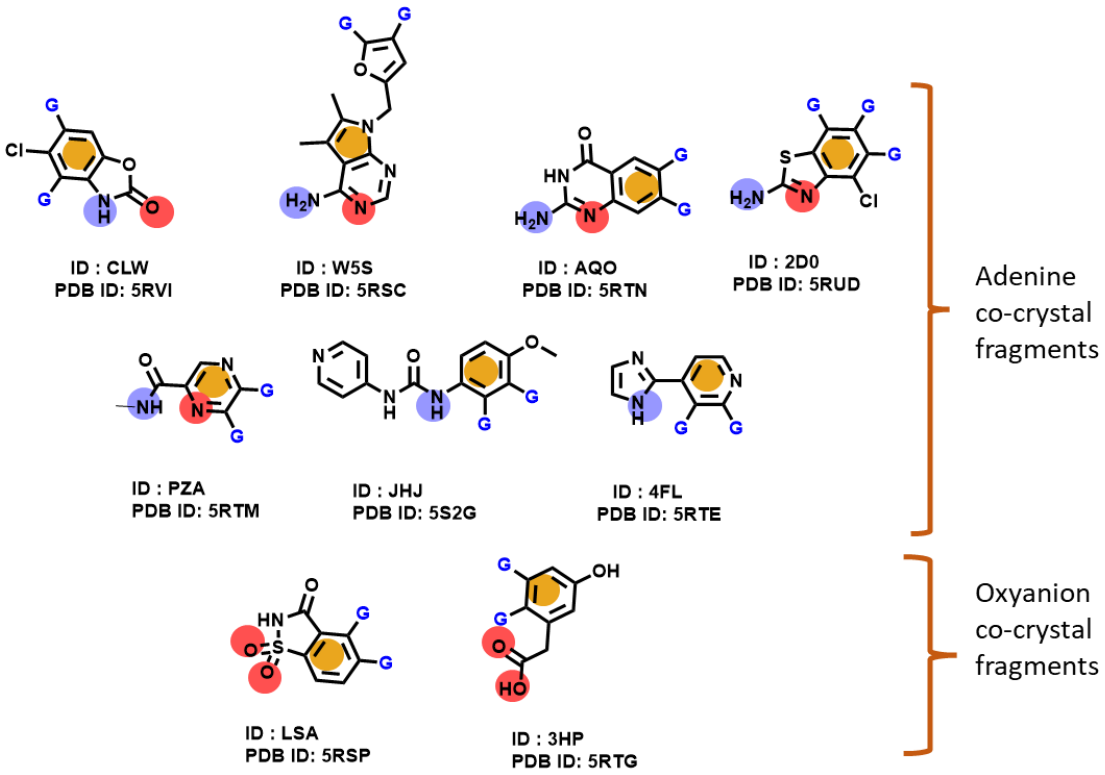

Figure 6: 2D structures of co-crystal fragments shortlisted for AI-driven fragment expansion approach and 2D-structures. Color-code: hydrogen bond acceptor features in red, hydrogen bond donor features in blue, and aromatic features in orange and sentinel token 'G' are assigned to each fragment and this token indicates the position for fragment growing into adjacent sub-site.

For the fragment growing experiment, we have selected the benzo[d]isothiazol-3(2H)-one 1,1-dioxide (LSA201), occupying the oxyanion sub-site of Mac1 (PDB ID of 5RSP) and anchoring it to the adenine sub-site. A close look at the binding mode revealed that the sulfonyl moiety of the fragment hydrogen bonding with Asp157 and the carbonyl moiety forms the hydrogen bonding with Gly130. As illustrated in figure 7, binding mode and orientation of the fragment suggest that 6th and 7th positions of the benzo[d]isothiazol-3(2H)-one 1,1-dioxide moiety are ideal for expansion. A conventional 'G' sentinel token will be incorporated into SMILES notation to indicate the direction in which the fragment will expand. We do this by inserting a 'G' token at the expansion site of the canonical SMILES string.

With each iteration, the LSTM adds a new atom or bond to the existing fragment, expanding it progressively. Over successive iterations, the LSTM gradually expands the fragment into a small molecule. For instance, from 'O=S2(=O)NC(=O)c1cccc(c12)G', it might predict next character like 'O=S2(=O)NC(=O)c1cccc(c12)**C**G', and in next step it will generate 'O=S2(=O)NC(=O)c1cccc(c12)**CC**G', and grows the molecule in accordance with the model's guidance. This process continues until it either encounters the End of Sequence (EOS) stop code or reaches the specified maximum length of the token sequence (max 74). We sampled 500 SMILES strings from 10 iterations from the pre-trained model and optimized by the fine-tuned model. 93% of the generated hits were valid and bearing the starting fragment. The fragment growing experiment was repeated twice for each of the 6^{th} and 7^{th} positions of the core.

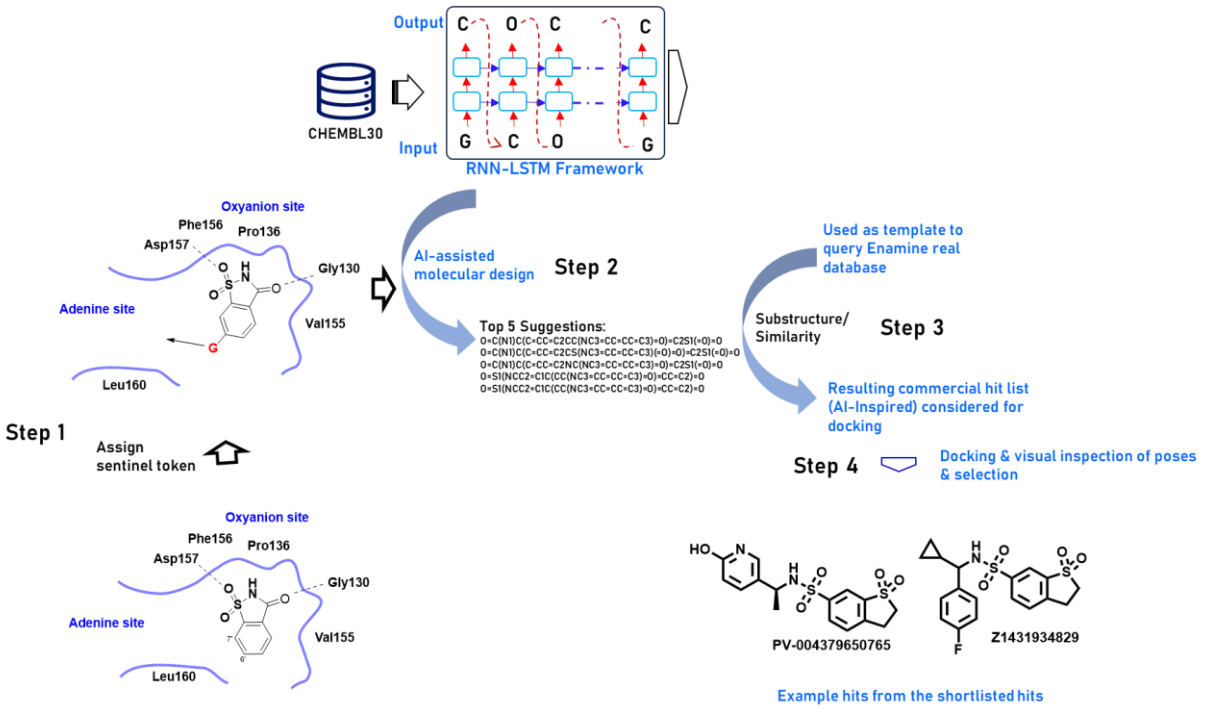

Figure 7: AI-driven fragment expansion using a starting co-crystal fragment occupying the oxyanion sub-site (PDB 5RSP). For a given co-crystal fragment, 6th and 7th positions of the core are suitable for expansion. For the 1st iteration, the sentinel token 'G' added to 6th position, which dictates the direction of expansion towards the adenine site, guiding the LSTM model. The fragment exit vector is represented as an arrow, providing insights into the extension process. This process will be repeated by defining sentinel token 'G' to the 7th position of the core, to generate ideas from that position. From top hits, the best two Enamine results are highlighted here.

A second example is pyrazine-2carboxamide (PZA), which occupied the adenine sub-site of Mac1 (PDB ID of 5RTM) and could expand into the oxyanion sub-site as shown in Figure 8. The binding mode indicated that the carboxamide moiety of this fragment forms hydrogen bonding with Asp22 and Val24, and the pyrimidine ring π-π stacks with Phe156. Based on the pose and orientation, we chose the 5th and 6th positions of the pyrimidine ring as suitable for further exploration. As mentioned in the above example, the SMILES of PZA is modified with a token 'G' added to specify the position (either 5th or 6th) to grow the fragment further. For each experiment, 500 suggestions were generated, and this process was repeated for each of the positions of interest explored.

We repeated these fragment growing experiments on each of the nine selected fragments. In each iteration, 500 SMILES were sampled using the pre-trained model then transfer learning techniques applied to refine the hits using fine-tuned model. Almost 90% of the generated hits retained the initial fragment and were valid SMILES. We conducted a visual analysis; we selected 5-10 molecules from each experiment.

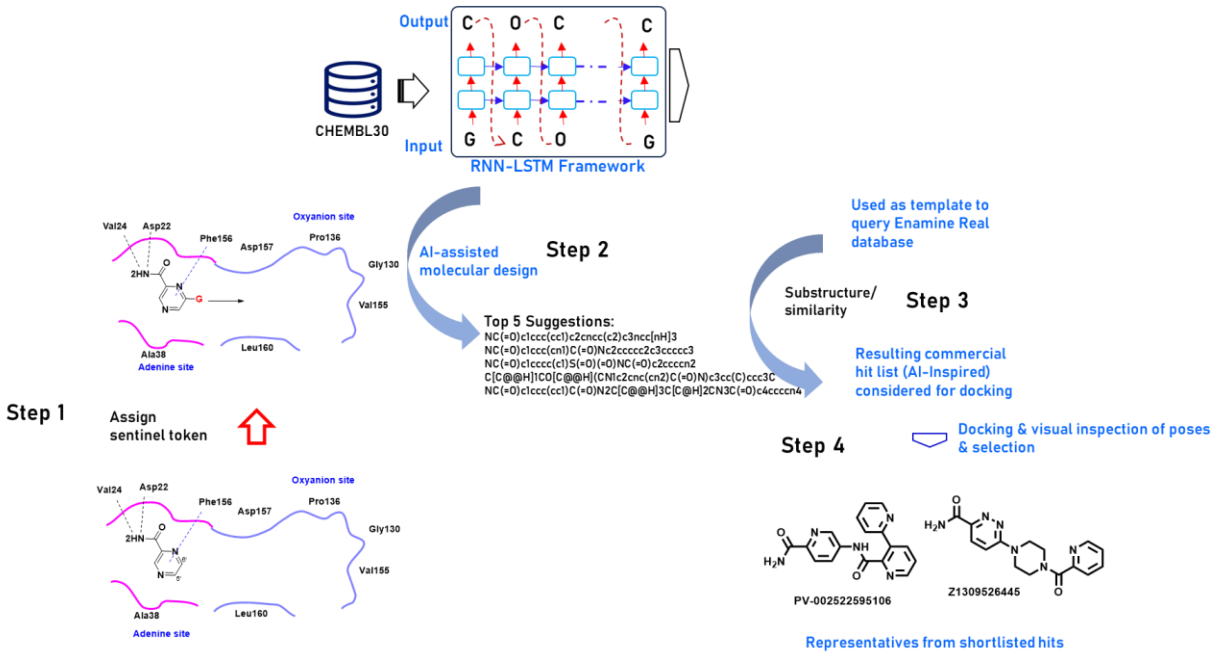

Figure 8: AI-driven fragment expansion using the co-crystal fragment occupied the adenine sub-site (PDB : 5RTM). For a given co-crystal fragment, 5th, and 6th positions of the core are suitable for expansion. For the 1st iteration, the sentinel token 'G' added to 6th position, which dictates the direction of expansion towards the adenine site, guiding the LSTM model. The fragment exit vector is represented as arrow, providing insights into the extension process. This process will be repeated by defining sentinel token 'G' to the 5th position of the core, to generate ideas from that position. From top hits, the best two Enamine results are highlighted here.

Because these molecules are new and not necessarily commercially available (or even stable and/or synthetically accessible), they served as templates to search the Enamine REAL database for structurally similar molecules. The 7,865 AI-inspired hits were then subjected to a wide range of filters (such as., Drug-like, PAINS, and toxic alerts), and cross-checked for duplicates with knowledge-based hit list.

3.3 Fragment linking using knowledge driven approach:

As part of our molecular design approach, our team also implemented knowledge-based fragment merging strategies. To select the appropriate fragments, we followed a set of criteria that included 1) identifying co-crystal fragments that are at a suitable distance from each other; 2) identifying linkers with the right geometry to connect them; and 3) ensuring the generated molecule maintains the interactions between the fragments and the protein.

Of the original twenty, six of the thirteen adenine co-crystal fragments and five of the seven oxyanion co-crystal fragments were selected for this study. Figure 9 illustrates the 2D-structures of the fragments that were selected for this study. Figure 10 illustrates the knowledge-based fragment merging process using quinolin-3-amine (4FS) which binds to Mac1's adenosine sub-site (PDB: 5RV4) and 3-hydroxy-2-methylbenzoic acid (PF0), which binds to Mac1's oxyanion sub-site (PDB ID of 5RTC). The superposition of these two crystal structures revealed that the 6th and 7th positions of the quinoline moiety of the 4FS is proximal to the benzoic acid moiety of the PF0, suggesting fragment merging. Using chemical intuition and expertise, we connected these fragments using common medicinal chemistry linkers and generated potential hits. These included commonly used linkers in medicinal chemistry such as amines, aliphatic linkers, carboxamides, and sulfonamides. Each fragment was explored for every possible combination and permutation to generate potential hits, using this strategy.

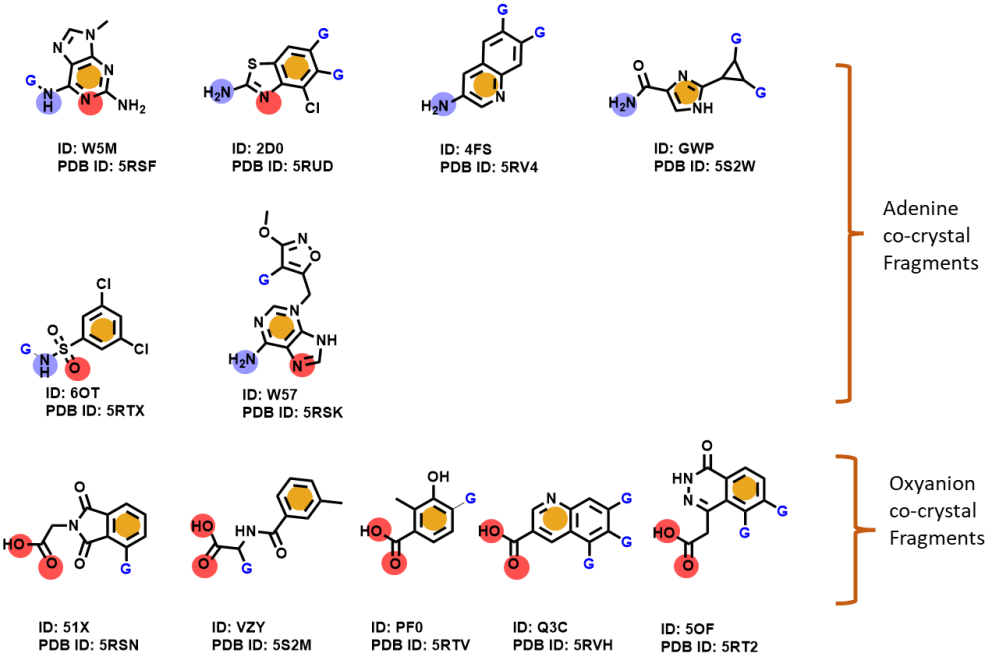

Figure 9: Co-crystallized fragments selected for knowledge-driven fragment merging approach. Notably, a majority of the shortlisted adenine co-crystal fragments contain an heteroaryl with one acceptor and one donor group in common. These are distinctly color-coded for clarity: Acceptor features in red, Donor features in blue, and aromatic features in orange.

The manually generated ideas (28) served as templates to query the Enamine REAL database, and the resulting hit list (9,135) were compiled with AI-inspired hits, remove duplicates, and resulting 12,800 compounds were processed and submitted to docking. Figure SF2 illustrates the process of selecting matching hits, referred to as AI-inspired hits, identified by using AI-generated hits as templates.

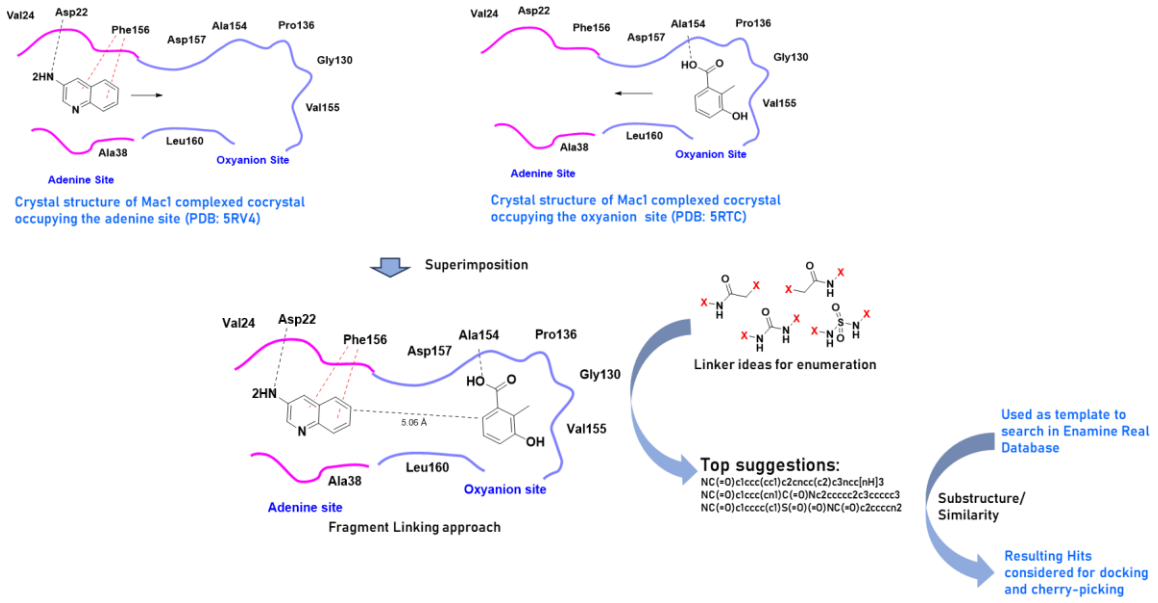

Figure 10: 2D-interaction diagram showed the selected crystal structures of Mac1 domain, one with complexed co-crystal with adenine site (PDB: 5RV4), and the other with oxyanion site (PDB: 5RTC). Another 2D-interaction diagram revealed the distances between both fragments as per the binding modes, interactions, and possible linking positions. Following this step, MedChem

linkers were used to connect both fragments, taking into account the positions in the fragment, the size of the linker, and the distance between them. The concepts generated from this process served as templates for our subsequent enamine search.

3.4 Hit list creation and processing:

In accordance with CACHE3 guidelines, we ensure our selected compounds from the Enamine REAL database met property cutoffs (logp </= 3.5 and mol weight < 400 da). Using our query templates (from the two approaches) for searching the Enamine REAL database, our strategy included structure searches for exact matches, substructure searches (where R groups were removed from molecules to find analogs or the same compound), and similarity searches with a 40% similarity threshold. The Enamine search found a total of 17,588 hits. Further processing was carried out to evaluate drug-like properties, and PAINS (Pan-Assay Interference Compounds) alerts. The resulting list of 12,800 compounds was further processed for docking studies.

3.5 Docking and compound selection:

For docking studies, we selected Mac1 crystal structure (PDB: 5SQJ), with a resolution of 1.05 Å and complexed with a small molecule inhibitor and top active Z5021668601 (with an IC_{50} of 0.5 μM). We have followed the same docking protocol as we used in the first two CACHE challenges.[34,35] We prepared the protein for docking with PREPARE and PROCESS—software within our drug discovery platform FORECASTER [26].

The library of 12,800 compounds were then processed with CONVERT (adding hydrogens, 2D to 3D conversion) and SMART (assignment of atom types, charges, etc.), all part of FORECASTER. We then split the library into 13 files to take advantage of our supercomputer parallelization and carried out the virtual screening with FITTED. Once complete, we collated the results and ranked the compounds by FITTED score. We then extracted the top 300 molecules for visual inspection.

3.6 Selection Process:

We classified the top 300 poses into different bins (“yes”, “no”, and “maybe”) according to several criteria: 1) overall fit of the ligands in the binding site, 2) occupancy of the correct sub-pockets and overlap with the relevant co-crystallized fragments, 3) relative ligand strain, and 4) number and type of fulfilled and unfulfilled interactions, e.g., hydrogen bonding, π-stacking, van der Waals interactions. It is important to note that this process is subjective and dependent on the experience (and biases) of each of the team members involved. While this may be seen as a limitation to the discovery of truly novel compounds, it has been shown to be an effective methodology.[36] To achieve a “yes” status, a compound must have passed a simple majority vote. For a “maybe”, the decision would have been split. Everything else was a “no”. Exceptionally, we would assign a “yes” if one member of our team decided to “champion” it. Each member of the team was allowed to champion a maximum of one compound.

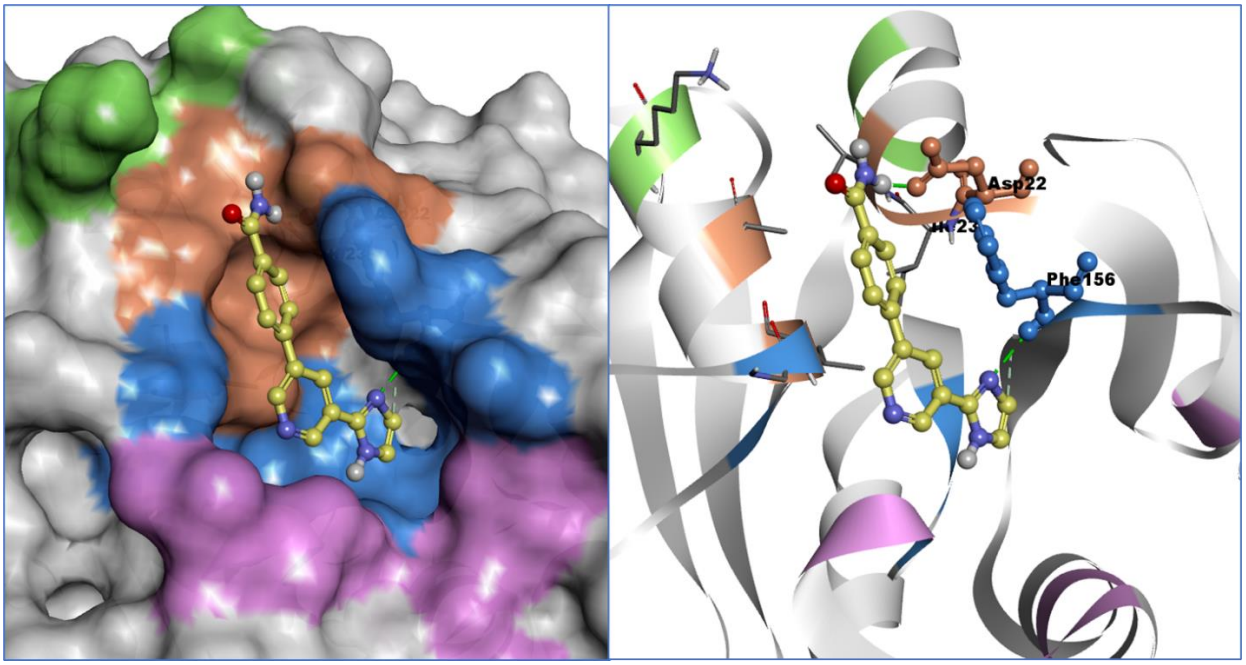

Figure 11: Left panel: predicted binding pose of PV-002833638168 in the crystal structure of the SARS-CoV-2 NSP3 macrodomain (PDB ID 5RTC), displayed in surface mode. Various sub-sites on the macrodomain surface are distinctly color-marked as described in Figure 4. Right panel: key interacting amino acids displayed as sticks, with the ligand (PV-002833638168) shown in ball-and-stick.

As an example, AI inspired hit PV-002833638168 (categorized in the "yes" bin) establishes key interactions (Figure 11). It retains the essential hydrogen bonding interactions with Asp22 from adenine site, and Phe156 from the oxyanion sub-site. Following this voting system, we selected 150 initial compounds (100 from our generative AI strategy and 50 from knowledge-based design). Following the CACHE3 guidelines, which include considerations like compound cost, availability, drug-like properties, PAINS alerts, delivery timelines, we refined our selection to 91 virtual hits, which were then forwarded for procuring and screening.

4. Summary and Outlook

In our pursuit to design competitive small molecule inhibitors for the Mac1 domain of NSP3, we embarked on two distinct approaches: AI-guided fragment expansion and knowledge-based fragment merging, specifically targeting the adenine and oxyanion sub-sites. These approaches purposefully led to molecules incorporating known binding fragments but not commercially available. These molecules were used as templates to identify similar molecules available for purchase. These analogues were then docked to Mac1 and through a consensus scoring methodology that incorporates the FITTED docking score, a visual score, and an interaction pattern, we identified 150 molecules, 91 of which were acquired for experimental assessment.

5. Methodology:

5.1 AI-guided Generative models

A recurrent neural network (RNN) is a type of neural network that is widely used for natural language processing tasks from simple language processing to complex cheminformatics problems. RNNs were used as autoencoders pre-trained to capture the structures (represented as SMILES strings) of 1,486,000 bioactive small molecules (covers around 14,850 biological targets) retrieved from ChEMBL30 (K_D, K_i, EC_{50}, IC_{50} < 1 µM).[25] This pre-trained model

comprises two LSTM layers, each possessing a hidden state vector of size 256, which is regularized using dropout (as shown in Figure 5). These LSTM layers lead to a dense output layer complemented by a softmax activation function. The LSTM layers receive input in the form of a one-shot encoded sequence derived from a molecule's SMILES string, segmented into tokens. Each SMILES string initiates with a 'G' token (representing "go") and concludes with an 'E' token, signifying the end of the string. Padding with an 'A' token equalizes SMILES sequence lengths for RNN input, enabling uniform batch processing and streamlining model training. Post-training, these tokens are discarded, permitting the RNN to generate variable-length molecular structures. During the character sampling process, we integrated an additional temperature parameter into the softmax function. This pre-trained model can generate random molecules that resemble the chemical space of CHEMBL, independent of any specific target or biological endpoint. Generated hits were rigorously evaluated and benchmarked using molecular sets (MOSES),[37] which assesses the quality (novelty, and validity) of generated SMILES. MOSES also checks for drug-likeness, synthetic feasibility, and compliance with medicinal chemistry standards, offering a comprehensive measure of the molecules' potential as drug candidates.

Our deep learning models were implemented using TensorFlow (v2.2, www.TensorFlow.org) in Python (v3.7, www.python.org). RDkit (www.rdkit.org) was extensively employed for cheminformatics activities, including property calculations, SMILES validation, molecular fingerprint calculations, and molecular clustering calculations. The Jupyter notebook (www.anaconda.org) facilitated a detailed analysis of the generated SMILES strings.

5.2 Rigid Docking

To dock molecules, Fitted (Flexibility Induced Through Targeted Evolutionary Description) uses a combination of matching algorithm and enhanced genetic algorithm.[38] The initial population of binding modes, which is optimized by the genetic algorithm, is generated using a matching algorithm. The latter is matching ligand pharmacophoric groups (e.g., phenyl = hydrophobic group, R-NH2 – hydrogen bond donor) with protein interaction sites, which are generated by PROCESS. For docking studies, shortlisted Mac1 crystal structure (PDB: 5SQJ) was downloaded from protein database using PREPARE module. PREPARE module cleans up the pdb files, then adds hydrogens, and optimizes their positions. For our study, we ignored the additional options to add particle waters and generate side chain conformations (for flexible docking studies). Prepared protein and ligand files are then saved as a .mol2 files and processed using PROCESS to assign protein atom types and charges. The ensemble of spheres cavity of the binding site was centered in the cavity and did not exceed 28 Å long. The grid resolution was 1.5 Å. Each ligand was charged using Sybyl Gasteiger-Huckel charges and processed using SMART. Large grids of spheres were prepared as well as constraint files. Default docking settings in FITTED were adjusted to increase the number of runs to three per molecule and three poses per molecules exported for subsequent analysis. PREPARE, SMART, and FITTED parameter files, as well as docking poses for the shortlisted 150 hits, are available for all users to replicate study and retrospective analyses.

6. Acknowledgments

The author(s) would like to express their sincere gratitude for the funding and support provided by the Minister of Innovation, Science and Industry and Conscience through CACHE Challenge, which has been crucial in facilitating this research. We are deeply thankful for their dedication to advancing open science research.

Les auteurs souhaitent exprimer leur sincère gratitude pour le financement et le soutien fournis par le ministre de l'Innovation, de la Science et de l'Industrie, ainsi que par Conscience à travers le défi CACHE, ces derniers ayant été cruciaux pour faciliter cette recherche. Nous sommes profondément reconnaissants pour leur dévouement à faire progresser la recherche scientifique ouverte.

6. References

(1) Wertheim, J. O.; Chu, D. K. W.; Peiris, J. S. M.; Kosakovsky Pond, S. L.; Poon, L. L. M. A Case for the Ancient Origin of Coronaviruses. *J. Virol.* **2013**, *87* (12), 7039–7045. https://doi.org/10.1128/JVI.03273-12.

(2) Dömling, A.; Gao, L. Chemistry and Biology of SARS-CoV-2. *Chem* **2020**, *6* (6), 1283–1295. https://doi.org/10.1016/j.chempr.2020.04.023.

(3) Tan, H.; Hu, Y.; Jadhav, P.; Tan, B.; Wang, J. Progress and Challenges in Targeting the SARS-CoV-2 Papain-like Protease. *J. Med. Chem.* **2022**, *65* (11), 7561–7580. https://doi.org/10.1021/acs.jmedchem.2c00303.

(4) Brian Chia, C. S.; Pheng Lim, S. A Patent Review on SARS Coronavirus Papain-Like Protease (PLpro) Inhibitors. *ChemMedChem* **2023**, *18* (16), e202300216. https://doi.org/10.1002/cmdc.202300216.

(5) Liu, Y.; Liang, C.; Xin, L.; Ren, X.; Tian, L.; Ju, X.; Li, H.; Wang, Y.; Zhao, Q.; Liu, H.; Cao, W.; Xie, X.; Zhang, D.; Wang, Y.; Jian, Y. The Development of Coronavirus 3C-Like Protease (3CLpro) Inhibitors from 2010 to 2020. *Eur. J. Med. Chem.* **2020**, *206*, 112711. https://doi.org/10.1016/j.ejmech.2020.112711.

(6) Chia, C. S. B.; Xu, W.; Shuyi Ng, P. A Patent Review on SARS Coronavirus Main Protease (3CLpro) Inhibitors. *ChemMedChem* **2022**, *17* (1), e202100576. https://doi.org/10.1002/cmdc.202100576.

(7) Chen, R.; Gao, Y.; Liu, H.; Li, H.; Chen, W.; Ma, J. Advances in Research on 3C-like Protease (3CLpro) Inhibitors against SARS-CoV-2 since 2020. *RSC Med. Chem.* **2023**, *14* (1), 9–21. https://doi.org/10.1039/d2md00344a.

(8) Li, X.; Song, Y. Structure and Function of SARS-CoV and SARS-CoV-2 Main Proteases and Their Inhibition: A Comprehensive Review. *Eur. J. Med. Chem.* **2023**, *260*, 115772. https://doi.org/10.1016/j.ejmech.2023.115772.

(9) Owen, D. R.; Allerton, C. M. N.; Anderson, A. S.; Aschenbrenner, L.; Avery, M.; Berritt, S.; Boras, B.; Cardin, R. D.; Carlo, A.; Coffman, K. J.; Dantonio, A.; Di, L.; Eng, H.; Ferre, R.; Gajiwala, K. S.; Gibson, S. A.; Greasley, S. E.; Hurst, B. L.; Kadar, E. P.; Kalgutkar, A. S.; Lee, J. C.; Lee, J.; Liu, W.; Mason, S. W.; Noell, S.; Novak, J. J.; Obach, R. S.; Ogilvie, K.; Patel, N. C.; Pettersson, M.; Rai, D. K.; Reese, M. R.; Sammons, M. F.; Sathish, J. G.; Singh, R. S. P.; Steppan, C. M.; Stewart, A. E.; Tuttle, J. B.; Updyke, L.; Verhoest, P. R.; Wei, L.; Yang, Q.; Zhu, Y. An Oral SARS-CoV-2 Mpro Inhibitor Clinical Candidate for the Treatment of COVID-19. *Science* **2021**, *374* (6575), 1586–1593. https://doi.org/10.1126/science.abl4784.

(10) Unoh, Y.; Uehara, S.; Nakahara, K.; Nobori, H.; Yamatsu, Y.; Yamamoto, S.; Maruyama, Y.; Taoda, Y.; Kasamatsu, K.; Suto, T.; Kouki, K.; Nakahashi, A.; Kawashima, S.; Sanaki, T.; Toba, S.; Uemura, K.; Mizutare, T.; Ando, S.; Sasaki, M.; Orba, Y.; Sawa, H.; Sato, A.; Sato, T.; Kato, T.; Tachibana, Y. 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. https://doi.org/10.1021/acs.jmedchem.2c00117.

(11) Shen, Z.; Ratia, K.; Cooper, L.; Kong, D.; Lee, H.; Kwon, Y.; Li, Y.; Alqarni, S.; Huang, F.; Dubrovskyi, O.; Rong, L.; Thatcher, G. R. J.; Xiong, R. Design of SARS-CoV-2 PLpro Inhibitors for COVID-19 Antiviral Therapy Leveraging Binding Cooperativity. *J. Med. Chem.* **2022**, *65* (4), 2940–2955. https://doi.org/10.1021/acs.jmedchem.1c01307.

(12) Rut, W.; Lv, Z.; Zmudzinski, M.; Patchett, S.; Nayak, D.; Snipas, S. J.; El Oualid, F.; Huang, T. T.; Bekes, M.; Drag, M.; Olsen, S. K. Activity Profiling and Crystal Structures of Inhibitor-Bound SARS-CoV-2 Papain-like Protease: A Framework for Anti–COVID-19 Drug Design. *Sci. Adv.* **2020**, *6* (42), eabd4596. https://doi.org/10.1126/sciadv.abd4596.

(13) Ip, J. D.; Chu, A. W.-H.; Chan, W.-M.; Leung, R. C.-Y.; Abdullah, S. M. U.; Sun, Y.; To, K. K.-W. Global Prevalence of SARS-CoV-2 3CL Protease Mutations Associated with Nirmatrelvir or Ensitrelvir Resistance. *eBioMedicine* **2023**, *91*. https://doi.org/10.1016/j.ebiom.2023.104559.

(14) Duan, Y.; Zhou, H.; Liu, X.; Iketani, S.; Lin, M.; Zhang, X.; Bian, Q.; Wang, H.; Sun, H.; Hong, S. J.; Culbertson, B.; Mohri, H.; Luck, M. I.; Zhu, Y.; Liu, X.; Lu, Y.; Yang, X.; Yang, K.; Sabo, Y.; Chavez, A.; Goff, S. P.; Rao, Z.; Ho, D. D.; Yang, H. Molecular Mechanisms of SARS-CoV-2 Resistance to Nirmatrelvir. *Nature* **2023**, *622* (7982), 376–382. https://doi.org/10.1038/s41586-023-06609-0.

(15) Kunze, F. A.; Hottiger, M. O. Regulating Immunity via ADP-Ribosylation: Therapeutic Implications and Beyond. *Trends Immunol.* **2019**, *40* (2), 159–173. https://doi.org/10.1016/j.it.2018.12.006.

(16) Schuller, M.; Correy, G. J.; Gahbauer, S.; Fearon, D.; Wu, T.; Díaz, R. E.; Young, I. D.; Carvalho Martins, L.; Smith, D. H.; Schulze-Gahmen, U.; Owens, T. W.; Deshpande, I.; Merz, G. E.; Thwin, A. C.; Biel, J. T.; Peters, J. K.; Moritz, M.; Herrera, N.; Kratochvil, H. T.; QCRG Structural Biology Consortium; Aimon, A.; Bennett, J. M.; Brandao Neto, J.; Cohen, A. E.; Dias, A.; Douangamath, A.; Dunnett, L.; Fedorov, O.; Ferla, M. P.; Fuchs, M. R.;

Gorrie-Stone, T. J.; Holton, J. M.; Johnson, M. G.; Krojer, T.; Meigs, G.; Powell, A. J.; Rack, J. G. M.; Rangel, V. L.; Russi, S.; Skyner, R. E.; Smith, C. A.; Soares, A. S.; Wierman, J. L.; Zhu, K.; O'Brien, P.; Jura, N.; Ashworth, A.; Irwin, J. J.; Thompson, M. C.; Gestwicki, J. E.; von Delft, F.; Shoichet, B. K.; Fraser, J. S.; Ahel, I. Fragment Binding to the Nsp3 Macrodomain of SARS-CoV-2 Identified through Crystallographic Screening and Computational Docking. *Sci. Adv.* **2021**, *7* (16), eabf8711. https://doi.org/10.1126/sciadv.abf8711.

(17) Rack, J. G. M.; Perina, D.; Ahel, I. Macrodomains: Structure, Function, Evolution, and Catalytic Activities. *Annu. Rev. Biochem.* **2016**, *85* (1), 431–454. https://doi.org/10.1146/annurev-biochem-060815-014935.

(18) Fehr, A. R.; Channappanavar, R.; Jankevicius, G.; Fett, C.; Zhao, J.; Athmer, J.; Meyerholz, D. K.; Ahel, I.; Perlman, S. The Conserved Coronavirus Macrodomain Promotes Virulence and Suppresses the Innate Immune Response during Severe Acute Respiratory Syndrome Coronavirus Infection. *mBio* **2016**, *7* (6), 10.1128/mbio.01721-16. https://doi.org/10.1128/mbio.01721-16.

(19) Gahbauer, S.; Correy, G. J.; Schuller, M.; Ferla, M. P.; Doruk, Y. U.; Rachman, M.; Wu, T.; Diolaiti, M.; Wang, S.; Neitz, R. J.; Fearon, D.; Radchenko, D. S.; Moroz, Y. S.; Irwin, J. J.; Renslo, A. R.; Taylor, J. C.; Gestwicki, J. E.; von Delft, F.; Ashworth, A.; Ahel, I.; Shoichet, B. K.; Fraser, J. S. Iterative Computational Design and Crystallographic Screening Identifies Potent Inhibitors Targeting the Nsp3 Macrodomain of SARS-CoV-2. *Proc. Natl. Acad. Sci.* **2023**, *120* (2), e2212931120. https://doi.org/10.1073/pnas.2212931120.

(20) Khoo, S. H.; Fitzgerald, R.; Fletcher, T.; Erwings, S. *Optimal dose and safety of molnupiravir in patients with early SARS-CoV-2: a Phase I, open-label, dose-escalating, randomized controlled study | Journal of Antimicrobial Chemotherapy | Oxford Academic*. 10.1093/jac/dkab318 (accessed 2023-12-13).

(21) Michael, S.; Heilbronner, R.; Lloyd, C. M.; Levitin, H. W. Paxlovid-Induced Tacrolimus Toxicity in the Treatment of COVID-19: A Case Report. *Cureus 15* (2), e35489. https://doi.org/10.7759/cureus.35489.

(22) Kirkwood-Johnson, L.; Marikawa, Y. Developmental Toxicity of Remdesivir, an anti-COVID-19 Drug, Is Implicated by in Vitro Assays Using Morphogenetic Embryoid Bodies of Mouse and Human Pluripotent Stem Cells. *Birth Defects Res.* **2023**, *115* (2), 224–239. https://doi.org/10.1002/bdr2.2111.

(23) Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: A Large-Scale Bioactivity Database for Drug Discovery. *Nucleic Acids Res.* **2012**, *40* (Database issue), D1100–D1107. https://doi.org/10.1093/nar/gkr777.

(24) Papadatos, G.; Davies, M.; Dedman, N.; Chambers, J.; Gaulton, A.; Siddle, J.; Koks, R.; Irvine, S. A.; Pettersson, J.; Goncharoff, N.; Hersey, A.; Overington, J. P. SureChEMBL: A Large-Scale, Chemically Annotated Patent Document Database. *Nucleic Acids Res.* **2016**, *44* (D1), D1220–D1228. https://doi.org/10.1093/nar/gkv1253.

(25) Zdrazil, B.; Felix, E.; Hunter, F.; Manners, E. J.; Blackshaw, J.; Corbett, S.; de Veij, M.; Ioannidis, H.; Lopez, D. M.; Mosquera, J. F.; Magarinos, M. P.; Bosc, N.; Arcila, R.; Kizilören, T.; Gaulton, A.; Bento, A. P.; Adasme, M. F.; Monecke, P.; Landrum, G. A.; Leach, A. R. The ChEMBL Database in 2023: A Drug Discovery Platform Spanning Multiple Bioactivity Data Types and Time Periods. *Nucleic Acids Res.* **2023**, gkad1004. https://doi.org/10.1093/nar/gkad1004.

(26) Therrien, E.; Englebienne, P.; Arrowsmith, A. G.; Mendoza-Sanchez, R.; Corbeil, C. R.; Weill, N.; Campagna-Slater, V.; Moitessier, N. Integrating Medicinal Chemistry, Organic/Combinatorial Chemistry, and Computational Chemistry for the Discovery of Selective Estrogen Receptor Modulatorswith FORECASTER, a Novel Platform for Drug Discovery. *J. Chem. Inf. Model.* **2011**, *52*, 210–224. https://doi.org/dx.doi.org/10.1021/ci2004779.

(27) Brosey, C. A.; Houl, J. H.; Katsonis, P.; Balapiti-Modarage, L. P. F.; Bommagani, S.; Arvai, A.; Moiani, D.; Bacolla, A.; Link, T.; Warden, L. S.; Lichtarge, O.; Jones, D. E.; Ahmed, Z.; Tainer, J. A. Targeting SARS-CoV-2 Nsp3 Macrodomain Structure with Insights from Human Poly(ADP-Ribose) Glycohydrolase (PARG) Structures with Inhibitors. *Prog. Biophys. Mol. Biol.* **2021**, *163*, 171–186. https://doi.org/10.1016/j.pbiomolbio.2021.02.002.

(28) Frick, D. N.; Virdi, R. S.; Vuksanovic, N.; Dahal, N.; Silvaggi, N. R. Molecular Basis for ADP-Ribose Binding to the Mac1 Domain of SARS-CoV-2 Nsp3. *Biochemistry* **2020**, *59* (28), 2608–2615. https://doi.org/10.1021/acs.biochem.0c00309.

(29) Suresh, N.; Kumar, N. C. A.; Subramanian, S.; Srinivasa, G. Memory Augmented Recurrent Neural Networks for De-Novo Drug Design. *PLOS ONE* **2022**, *17* (6), e0269461. https://doi.org/10.1371/journal.pone.0269461.

(30) Segler, M. H. S.; Kogej, T.; Tyrchan, C.; Waller, M. P. Generating Focused Molecule Libraries for Drug Discovery with Recurrent Neural Networks. *ACS Cent. Sci.* **2018**, *4* (1), 120–131. https://doi.org/10.1021/acscentsci.7b00512.

(31) Yasonik, J. Multiobjective de Novo Drug Design with Recurrent Neural Networks and Nondominated Sorting. *J. Cheminformatics* **2020**, *12* (1), 14. https://doi.org/10.1186/s13321-020-00419-6.
(32) Jang, S. H.; Sivakumar, D.; Mudedla, S. K.; Choi, J.; Lee, S.; Jeon, M.; Bvs, S. K.; Hwang, J.; Kang, M.; Shin, E. G.; Lee, K. M.; Jung, K.-Y.; Kim, J.-S.; Wu, S. PCW-A1001, AI-Assisted de Novo Design Approach to Design a Selective Inhibitor for FLT-3(D835Y) in Acute Myeloid Leukemia. *Front. Mol. Biosci.* **2022**, *9*, 1072028. https://doi.org/10.3389/fmolb.2022.1072028.
(33) Zhavoronkov, A.; Ivanenkov, Y. A.; Aliper, A.; Veselov, M. S.; Aladinskiy, V. A.; Aladinskaya, A. V.; Terentiev, V. A.; Polykovskiy, D. A.; Kuznetsov, M. D.; Asadulaev, A.; Volkov, Y.; Zholus, A.; Shayakhmetov, R. R.; Zhebrak, A.; Minaeva, L. I.; Zagribelnyy, B. A.; Lee, L. H.; Soll, R.; Madge, D.; Xing, L.; Guo, T.; Aspuru-Guzik, A. Deep Learning Enables Rapid Identification of Potent DDR1 Kinase Inhibitors. *Nat. Biotechnol.* **2019**, *37* (9), 1038–1040. https://doi.org/10.1038/s41587-019-0224-x.
(34) Burai-Patrascu, M.; Nivedha, A. K.; Rostaing, O.; Chukka, P.; Moitessier, N.; Pottel, J. The First CACHE Challenge–Identifying Binders of the WD-Repeat Domain of Leucine-Rich Repeat Kinase 2. *ChemRxiv* **2022**.
(35) Nivedha, A. K.; Burai-Patrascu, M.; Rostaing, O.; Chukka, P.; Singh, B.; Janezic, M.; Moitessier, N.; Pottel, J. *The Second CACHE Challenge - Targeting the RNA-Binding Pocket of the SARS-CoV2 Nonstructural Protein 13 via a consensus-scoring method and FITTED templated docking*. 10.26434/chemrxiv-2023-5g38r (accessed 2023-12-13).
(36) Irwin, J. J.; Shoichet, B. K. Docking Screens for Novel Ligands Conferring New Biology. *J. Med. Chem.* **2016**, *59* (9), 4103–4120. https://doi.org/10.1021/acs.jmedchem.5b02008.
(37) Polykovskiy, D.; Zhebrak, A.; Sanchez-Lengeling, B.; Golovanov, S.; Tatanov, O.; Belyaev, S.; Kurbanov, R.; Artamonov, A.; Aladinskiy, V.; Veselov, M.; Kadurin, A.; Johansson, S.; Chen, H.; Nikolenko, S.; Aspuru-Guzik, A.; Zhavoronkov, A. Molecular Sets (MOSES): A Benchmarking Platform for Molecular Generation Models. *Front. Pharmacol.* **2020**, *11*, 565644. https://doi.org/10.3389/fphar.2020.565644.
(38) Corbeil, C. R.; Englebienne, P.; Moitessier, N. Docking Ligands into Flexible and Solvated Macromolecules. 1. Development and Validation of FITTED 1.0. *J. Chem. Inf. Model.* **2007**, *47*, 435–449. https://doi.org/10.1021/ci6002637.

Supporting files

- Protein files used for docking.
- SMARTed library files for AI-assisted molecular design, knowledge-based fragment designs (mol2 format).
- Data analysis file, which includes the initial 150 selected compounds with scores for docking (Excel format).
- Visualized docked poses for each site (sdf format) for the initial selection of 150 compounds.
- Mac1 known inhibitors with computed properties in Excel format.

ORCID:

Suneel Kumar B.V.S: 0000-0001-6782-6446
Mihai Burai-Patrascu: 0000-0001-9289-7887
Ophélie Rostaing: 0000-0003-3921-2978
Matej Janezic: 0000-0003-3155-296X
Antoine Moitessier: 0009-0002-9255-6614
Nicolas Moitessier: 0000-0001-6933-2079
Joshua Pottel: 0000-0002-4455-4713