On the generation of novel ligands for SARS-CoV-2 protease and ACE2 receptor via constrained graph variational autoencoders

Jasper Kyle Catapang[1,*] **and Junie B. Billones**[2]

[1,2]Department of Physical Sciences and Mathematics, University of the Philippines Manila, Manila City, Philippines

[*]e-mail: jcatapang@up.edu.ph

ABSTRACT

SARS-CoV-2 has no known vaccine nor any effective treatment that has been released for clinical trials yet. This has ultimately paved the way for novel drug discovery approaches since although there are multiple efforts focused on drug repurposing of clinically-approved drugs for SARS-CoV-2, it is also worth considering that these existing drugs can be surpassed in effectivity by novel ones. This research focuses on the generation of novel candidate inhibitors via constrained graph variational autoencoders and the calculation of their Tanimoto similarities against existing drugs—repurposing these existing drugs and considering the novel ligands as possible SARS-CoV-2 main protease inhibitors and ACE2 receptor blockers by docking them through PyRx and ranking these ligands.

Introduction

Coronaviruses are members of a virus family called Coronaviridae[1]. Severe acute respiratory syndrome coronavirus (SARS-CoV) and the Middle East Respiratory Syndrome (MERS) virus revealed how coronaviruses can be a source of critically serious human tract infections. The first case of confirmed SARS-CoV was found in China, in November 2002. New cases of SARS emerged in China in February 2003. As for MERS-CoV, the first case occurred in June 2012 in Saudi Arabia[2]. These events have shown that the coronaviruses must not be underestimated. These have also highlighted the significance of advancing the knowledge on the replication of such viruses[3].

A typical pneumonia case emerged in Wuhan, China in December 2019. It was identified to have been caused by a completely unique coronavirus, named 2019 novel coronavirus (2019-nCoV), now called COVID-19. Multiple tests on COVID-19 patients have shown that the infection caused clusters of severe respiratory disorder just like SARS-CoV[4]. In the early stages, 2019-nCoV did not transmit between people. Newer epidemiological data however suggest that the new virus has undergone

human host adaptation[5]. SARS-CoV-2 has become more efficient in human to human transmission. SARS-CoV-2 is the seventh member of the family of coronaviruses that infect humans. SARS-CoV-2 enters target cells through an endosomal pathway and also uses the identical cell entry receptor, Angiotensin-converting enzyme II (ACE2), similar to SARS-CoV[5,6].

Variational autoencoders demonstrated that they are efficient generators of graphs, and by extension, molecular graphs of compounds[7]. They provide a stable and effective approach to construct new candidates for SARS-CoV-2 protease inhibitors. Atoms can be represented as nodes in a molecular graph. The encoder converts a molecular graph into a continuous latent representation, and the decoder converts these latent representations back to molecular graphs[7]. This process successfully generates new molecules.

The following are the researchers' contributions:

1. Generation of novel compounds that may be candidate inhibitors for SARS-CoV 2 main protease through constrained graph variational autoencoders
2. Similarity search of generated novel compounds to existing drugs for repurposing and comparison of binding energies
3. Molecular docking of the repurposed drugs and novel ligands that will be used for experimental assays

The research paper is divided into five parts. In the next section, the related works are discussed to show the background and different approaches of drug design. In the methodology section, the step-by-step process of creating the entire constrained graph variational autoencoder (CGVAE) for generation of novel ligands is discussed—from gathering the dataset to obtaining SMILES format outputs. The calculation of the Tanimoto coefficients for drug repurposing would also be included. In the fourth section, the results of the experimental study are shown and explained. Then, in the fifth section, the conclusion of the study is given with the statistics and findings by the researchers.

Related Work

Probabilistic generative molecular design

Molecules are microscopic compounds with more or less well-defined structure. A molecular structure according to[8] distinguishes a molecule from a collection of its constituent atoms. Electronic structure is closely related to molecular structure. The attracting forces exerted on electrons by the nuclei give order to the distribution of electrons within a molecule. The molecular structure is determined by three elements: constitution, configuration, and conformation[9]. Constitution is related to the sequence of chemical bonding of atoms and is expressed by topological descriptors, presence, and also by absence of functional groups[7]. Configuration, on the other hand, is defined by a 3D spatial arrangement of atoms[7]. It is characterized by the valence angles of all atoms that are directly linked to at least two other atoms. Configuration is expressed by shape descriptors[7]. Lastly, the

conformation of a molecule represents one of multiple thermodynamically stable three-dimensional spatial arrangements of its atoms. Researchers focus on constitution because it has the most contribution on molecular properties and activities.
Molecular representations provide machine-readable representations of molecular structure. One molecular structure can have many valid and unambiguous molecular representations. The most common molecular representations are line notations and molecular graphs. Line notations represent molecular structure as a linear string of characters while molecular graphs represent molecular structure as a graph. Line notations include the Simplified Molecular-Input Line-Entry System (SMILES) and the IUPAC Chemical Identifier (InChI). For the purpose of this paper, the SMILES representation would be used since this is the most used in deep learning.

Simplified Molecular-Input Line-Entry System

SMILES includes specifications of the following elements of a molecular structure: atoms, bonds, branching, rings, disconnections, isomerism, and reactions[10]. A common approach used in probabilistic generative molecular design (PGMD) is to train a generative model on SMILES. This model is used to generate SMILES for new molecules with a desired property or activity. There is no guarantee that the generated SMILES would even be valid or if they will be able to represent a reasonable molecular structure. The validity problem is broken into two: SMILES's semantics and syntax[9].

DeepSMILES is an adaptation of SMILES for use in machine learning of chemical structures[11]. DeepSMILES solves two primary invalid syntax causes. Its syntax avoids unbalanced parentheses by only using close parentheses, where the number of parentheses indicates the branch length. It also avoids the problem of pairing ring closure symbols by using only a single symbol at the ring closing location, where the symbol indicates the ring size[11].

Graph convolutional network components

Molecular descriptors[12] are numbers that capture particular features of molecular structure. Molecular descriptors can be obtained from molecular graphs via matrices and graph enumeration. A graph convolutional network (GCN)[13] are utilized for discovering functional groups in organic molecules that contribute to specific molecular properties[14]. Attributes of latent representations and latent space learned in a deep generative model (DGM) are determined by the latent variable model (LVM) used in the DGM, the sample data representations, and the DGM architecture, and DGM algorithms.

PGMD Architecture

The PGMD architecture borrows from the variational autoencoder architecture. The diagram in Figure 1 illustrates the architecture.

PGMD-VAE approaches

There are two general types of PGMD-VAE approaches: SMILES-based and molecular graph-based. Some examples of SMILES-based PGMD are CVAE, GVAE, and SD-VAE. CVAE[7] uses single-layer Long-Short Term Memory (LSTM) RNNs for both the encoder and the decoder. Latent representations obtained using the CVAE contain complete sequences and can be used to produce coherent new sequences that interpolate between known parts. GVAE[7] encodes and decodes directly from and to parse trees derived from SMILES, making sure that the generated outputs are always valid based on the grammar. As a consequence of producing outputs which are both syntactically true and semantically consistent, the SD-VAE allows for more improvement than the GVAE. The molecular latent space learned is more discriminative.

For molecular graph-based VAE approaches, one well-known example is the constrained graph variational autoencoder or CGVAE. CGVAE[15] is a sequential generative model for molecular graphs built in encoder and decoder from a VAE with Gated Graph Neural Networks (GGNNs)[13]. Instead of whole molecules it studies latent representations of the atoms referred. Additionally the decoder shapes nodes and edges. Decoding is achieved by initializing a set of potential nodes to link first[15]. The decoder then iterates through the specified edges, performs an edge selection and edge marking step for the currently oriented node, transfers the new molecular graph attached to a GGNN to update the node representations, and continues this procedure until an edge to a special stop node is chosen. For a new node in the current linked network, this whole cycle is replicated, and terminates if there are no suitable candidates. The decoder uses a valence mask to help ensure accurate molecule formation and avoid the creation of additional bonds on atoms that have already been allocated the permitted number of bonds for that particular type of atom[15].

Molecular docking

Molecular docking is a procedure for calculating the interaction between biomolecules like drugs and receptors for drug design and discovery. Docking is the method of attempting to place a ligand into an appropriate binding site of a target receptor with noncovalent bonding in order to create a specific, potential, and stable compound[16]. The ligand-protein interaction is determined by a lower binding free energy through the use of various scoring functions[17] such as force field-based[18], empirical[19], or knowledge-based scoring functions[20].

Software name	Function
AutoDock	protein-ligand docking
AutoDock Vina	molecular docking and virtual screening
Glide	virtual screening and high throughput ligand-receptor
HADDOCK	protein-protein docking

Table 1. Several well-known docking softwares

Enumerated in Table 1 are some of the commonly-used docking softwares. AutoDock[21] performs the docking of the flexible ligand to a fixed set of grids on a target. AutoDock Vina[22] is more accurate and faster than AutoDock. In Glide[23], the best binding molecules are determined by Monte Carlo sampling. HADDOCK[24] predicts docking based on the bio-physiochemical interaction data like chemical shift perturbation.

Tanimoto similarity

The Tanimoto similarity is the most common similarity metric for molecules. Consider fingerprint sets A and B of molecules A and B. AB is the set of common bits of fingerprints of both molecule A and B. The Tanimoto coefficient ranges from 0 when the fingerprints have no bits in common, to 1 when the fingerprints are identical[25]. The formula for the Tanimoto coefficient using sets A and B is: $T(A,B) = \frac{(A \wedge B)}{(A + B - A \wedge B)}$.

Methodology

The research can be split into four modules: (1) fetching and preprocessing of the inhibitor dataset, (2) ligand generation, (3) Tanimoto similarity on existing drugs, and (4) docking and ranking. With the related works discussed, this experimental study would utilize CGVAE[15] and PyRx as docking software.

Inhibitor dataset preparation

The research have two batches. The first batch targets the viral protease, while the other one targets the cell entry receptor. This decision provides a fallback if one of those targets fail to show desirable results.

For the first batch, the researchers are targeting the 6LU7 SARS-COV-2 main protease. The datasets contain around 490 TRPM8 inhibitors and 14 HIV inhibitors. They were both obtained from ChEMBL through diversity reduction among millions of inhibitors. ChEMBL is a manually curated database of bioactive molecules with drug-like properties. It brings together chemical, bioactivity and genomic data to aid the translation of genomic information into effective new drugs. The datasets

contain SMILES formatted inhibitors. Each inhibitor was calculated of its quantitative estimation of drug-likeness (QED) through the Python package rdkit. The QED is useful for addressing the problem of molecular target druggability assessment.

SMILES representation
CC(C)[C@@H]1CC[C@@H](C)C[C@H]1NC(=O)c2ccc(C)cc2
CC(C)[C@@H]1CC[C@@H](C)C[C@H]1NC(=O)c2ccc(cc2)c3ccccc3
CC(C)[C@@H]1CC[C@@H](C)C[C@H]1NC(=O)CCCCCCc2ccccc2
CC(C)[C@@H]1CC[C@@H](C)C[C@H]1NC(=O)CCCCCCCc2ccccc2
CC(C)[C@@H]1CC[C@@H](C)C[C@H]1NC(=O)c2cccc(Cl)c2
CC(C)[C@@H]1CC[C@@H](C)C[C@H]1NC(=O)c2ccc(Cl)cc2
COc1ccc(cc1)C(=O)N[C@@H]2C[C@H](C)CC[C@H]2C(C)C
CC(C)[C@@H]1CC[C@@H](C)C[C@H]1NC(=O)Oc2ccc(C)cc2
CC(C)[C@@H]1CC[C@@H](C)C[C@H]1NC(=O)Oc2cccc(c2)C(F)(F)F
CC(C)[C@@H]1CC[C@@H](C)C[C@H]1NC(=O)Oc2ccc(cc2)C(F)(F)F

Table 2. Several SMILES representation of data samples obtained from ChEMBL

Illustrated in Table 2 are some of the data samples obtained from the ChEMBL database in SMILES format.

Ligand generation

The combined dataset containing an estimate of 500 inhibitors served as input for the CGVAE as suggested in[15]. A dense GGNN model with the following parameters was used on the dataset: clamp gradient norm of 1.0, a QED trade-off lambda of 10, prior learning rate of 0.05, 3 epochs, hidden layer size of 20, a learning rate of 0.001, and a breadth-first search (BFS) path count of 30. These parameters were fine-tuned via trial and error as grid-search would have taken too long to find the optimal parameter values. The CGVAE model produced a .smi file as an output, along with 2D representations saved as .png files via rdkit.

Similarity search

The generated ligands are compared against existing drugs from DrugBank via Tanimoto similarity as previously discussed[25]. The produced comparisons were sorted in descending order and saved as a .csv file with query SMILES, target SMILES, and Tanimoto coefficients as features.

Docking and ranking

Other reported drugs that are being tested for effectivity against SARS-CoV-2 have also been added for docking. Using PyRx with AutoDock Vina as the backend, the novel ligands and repurposed drugs were docked to the targets: 6LU7 main protease and 1R42 ACE2 receptor, and their binding affinities in kcal/mol were calculated. The lower the binding energy, the greater the binding affinity is. The results were saved in a .xlsx file.

Results and Discussion

Ligand generation

The CGVAE managed to generate 100 novel candidate ligands for the SARS-CoV-2 main protease and the Angiotensin-converting enzyme II receptor. Figure 2 illustrates M0043, a novel candidate ligand that was generated via CGVAE for the SARS-CoV-2 main protease.

Similar drugs

By calculating the most similar drugs in DrugBank via their Tanimoto coefficients, the researchers were able to find three most plausible drugs. The novel ligands are labeled MX where X is the Xth generated ligand.

Query	Target	Tanimoto coefficient
M0053	TMC-310911	0.653
M0083	TMC-310911	0.648
M0053	Tipranavir	0.628
M0083	Tipranavir	0.619
M0090	TMC-310911	0.612
M0081	TMC-310911	0.604

Table 3. Tanimoto coefficients: Novel drugs vs. DrugBank drugs

Query	Target	Tanimoto coefficient
TMC-310911	Darunavir	0.768
TMC-310911	Brecanavir	0.735

Table 4. Tanimoto coefficients: TMC-310911 vs. DrugBank drugs

TMC-310911 and tipranavir are the most similar to the novel drugs generated by CGVAE as shown on Table 3. It is also worth mentioning that TMC-310911 is very similar to darunavir and brecanavir as shown on Table 4—but brecanavir was discontinued. TMC-310911 is a new investigational protease inhibitor that is structurally similar darunavir. It is being investigated for use in HIV-1 infections. Tipranavir is a sulfonamide-containing dyhydropyrone and a nonpeptidic protease inhibitor that targets the HIV protease. Darunavir is a protease inhibitor used with other HIV protease inhibitor drugs as well as ritonavir for the effective management of HIV-1 infection.

Docking and ranking

SARS-CoV-2 main protease

PDB ID: 6LU7

Active site: x = -11.70, y = 13.90, z = 70.55 (r = 12)

Ligand	Binding affinity (kcal/mol)
M0043	-8.3
M0074	-8.1
M0093	-7.8
Darunavir	-6.8
Tipranavir	-6.6
N3 (inhibitor)	-4.8
Favipiravir	-4.7

Table 5. Binding affinities of novel ligands and existing drugs on the active site of SARS-CoV-2 main protease

Out of 100 novel generated compounds, only three compounds have significantly lower binding energies when docked on 6LU7. Table 5 shows that the novel ligands M004, M0074, and M0093 have lower binding affinities: -8.3 kcal/mol, -8.1 kcal/mol, and -7.8 kcal/mol respectively, than darunavir, tipranavir, N3 inhibitor, and favipiravir with binding affinities in kcal/mol: -6.8, -6.6, -4.8, and -4.7, respectively.

ACE2 receptor

PDB ID: 1R42

Active Site: x = 52.59, y = 63.27, 26.46 (r=12)

Ligand	Binding affinity (kcal/mol)
M0076	-8.4
M0081	-6.9
M0021	-6.9
M0035	-6.7
MLN 4760	-4.7

Table 6. Binding affinities of novel ligands and an inhibitor named MLN 4760 on the active site of Angiotensin converting enzyme-related carboxypeptidase

Out of 100 novel generated compounds, there are four compounds that have significantly lower binding energies when docked on 14R2. M0076 (-8.4 kcal/mol), M0081 (-6.9 kcal/mol), M0021 (-6.9 kcal/mol), M0035 (-6.7 kcal/mol) have lower binding affinities than the inhibitor MLN 4760 with -4.7 kcal/mol, according to Table 6.

Figure 3 illustrates the chemical structures of the novel ligands for the SARS-Cov-2 main protease: M0043, M0074, and M0093. Figure 4 shows the chemical structures of the novel ligands—M0076, M0081, M0021, and M0035— for ACE2 receptor, where SARS-CoV-2 binds to.

Figures 5-7 show the interaction diagrams of M0043, M0074, and M0093, the novel ligands for SARS-CoV-2 main protease, meanwhile, Figures 8-11 show the interaction diagrams of M0076, M0081, M0021, and M0035, the novel ligands for the ACE2 receptor.

Conclusion

This research uses the ChEMBL database to extract around 500 known diverse drugs as input for a constrained graph variational autoencoder in order to generate 100 novel ligands. These novel ligands were compared against existing drugs in DrugBank by obtaining their pairwise Tanimoto coefficients.

The DrugBank drugs most similar with the different novel drugs were found to be TMC-310911 and tipranavir with Tanimoto coefficients greater than 60%. TMC-310911 is also structurally very similar to darunavir. These three existing drugs, along with other reported drugs used for SARS-CoV-2 testing, and the novel ligands were docked and ranked using PyRx with AutoDock Vina as backend.

The novel ligands for 6LU7 (main protease of SARS-CoV2)—M004, M0074, and M0093—have lower binding affinities: -8.3 kcal/mol, -8.1 kcal/mol, and -7.8 kcal/mol respectively, than darunavir, tipranavir, N3 inhibitor, and favipiravir with binding affinities in kcal/mol: -6.8, -6.6, -4.8, and -4.7, respectively.

The novel ligands for 14R2 (ACE2)—M0076 (-8.4 kcal/mol), M0081 (-6.9 kcal/mol), M0021 (-6.9 kcal/mol), M0035 (-6.7 kcal/mol)—have lower binding affinities than the inhibitor MLN 4760 with -4.7 kcal/mol.

CGVAE has managed to generate three new 6LU7 ligands and four new 14R2 ligands.

Acknowledgements

The authors would like to thank Geoffrey Solano, Kevin Sison, and Miguel Inco for their invaluable insights and suggestions.

References

1. Song; Z. et al.; "From SARS to MERS; Thrusting Coronaviruses into the Spotlight". Viruses 11, **2019**.

2. de Wit; E.; van Doremalen; N.; Falzarano; D. and Munster; V. J.; "SARS and MERS: recent insights into emerging coronaviruses". Nat. Rev. Microbiol. 14, **2016**; 523–534.

3. Menachery; V. D. et al.; "A SARS-like cluster of circulating bat coronaviruses shows potential for human emergence". Nat. Med. 21, **2015**; 1508–1513.

4. Huang; C. et al.; "Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China". Lancet, **2020**. doi:10.1016/S0140-6736(20)30183-5.

5. Zhou; P. et al.; "Discovery of a novel coronavirus associated with the recent pneumonia outbreak in humans and its potential bat origin". Microbiology 104, **2020**.

6. Letko; M. and Munster; V.; "Functional assessment of cell entry and receptor usage for lineage B β-coronaviruses, including 2019-nCoV". Microbiology 6117, **2020**.

7. Chang; D.; “Concept-Oriented Deep Learning: Generative Concept Representations". arXiv preprint **2018**; arXiv:1811.06622.

8. Sukumar; N.; “Molecular Similarity and Molecule Structure”. ISPC, **2007**; San Francisco.

9. Nikolova; N. and Jaworska; J.; “Approaches to Measure Chemical Similarity: A Review”. QSAR and Combinatorial Science 22 **2003**; 1006–1026.

10. Weininger D.; “SMILES; A Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules”. J Chem Inf Comput Sci 28:31–36, **1988**.

11. O'Boyle; N.M. and Dalke; A.; "DeepSMILES: An Adaptation of SMILES for Use in Machine-Learning of Chemical Structures". ChemRxiv, **2018**.

12. Grisoni; F.; "Molecular Descriptors. Theory and Tips for Real-world Applications." ETH Zurich, **2017**.

13. Wu; Z.; Pan; S.; Chen; F.; Long; G.; Zhang; C.; and Yu; P.S.; "A Comprehensive Survey on Graph Neural Networks". arXiv preprint **2019**; arXiv:1901.00596.

14. Pope; P.; Kolouri; S.; Rostrami; M.; Martin; C.; and Hoffmann; H.; "Discovering Molecular Functional Groups Using Graph Convolutional Neural Networks". arXiv preprint **2018**; arXiv:1812.00265.

15. Liu; Q.; Allamanis; M.; Brockschmidt; M.; and Gaunt; A.L.; "Constrained Graph Variational Autoencoders for Molecule Design". Neural Information Processing Systems (NIPS), **2018**.

16. Dar; A.M.; and Mir; S.; "Molecular Docking: Approaches; Types; Applications and Basic Challenges". J Anal Bioanal Tech 8 **2017**: 2.

17. Kitchen; D.B.; Decornez; H.; Furr; J.R.; and Bajorath; J.; "Docking and Scoring in Virtual Screening for Drug Discovery: Methods and Applications". Nature reviews Drug discovery 3(11) **2004**: 935.

18. Ewing; T.J.A.; Makino; S.; Skillman A.G.; and Kuntz; I.D.; "DOCK 4.0: Search Strategies for Automated Molecular Docking of Flexible Molecule Databases". Journal of computer-aided molecular design 15(5) **2001**: 411–28.

19. Wang; R.; Lai; L.; and Wang; S.; "Further Development and Validation of Empirical Scoring Functions for Structure-Based Binding Affinity Prediction". Journal of computer-aided molecular design 16(1) **2002**: 11–26.

20. Gohlke; H.; Hendlich; M.; and Klebe; G.; "Knowledge-Based Scoring Function to Predict Protein-Ligand Interactions 1." Journal of molecular biology 295(2) **2000**: 337–56.

21. Morris; G.M.; et al.; "AutoDock4 and AutoDockTools4: Automated Docking with Selective Receptor Flexibility". Journal of computational chemistry 30(16) **2009**: 2785–91.

22. Trott; O.; and Olson; A.J.; "AutoDock Vina: Improving the Speed and Accuracy of Docking with a New Scoring Function, Efficient Optimization, and Multithreading". Journal of computational chemistry 31(2) **2010**: 455–61.

23. Friesner; R.A.. et al.; "Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy". Journal of Medicinal Chemistry 47(7) **2004**: 1739–49.

24. Dominguez; C.; Boelens; R.; and Bonvin; A.M.J.J; "HADDOCK: A Protein- Protein Docking Approach Based on Biochemical or Biophysical Information". Journal of the American Chemical Society 125(7) **2003**: 1731–37.

25. Bajusz; D.; Racz; A.; and Heberger; K.; "Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations?". Journal of Cheminformatics volume 7, Article number 20, **2015**.

Figures

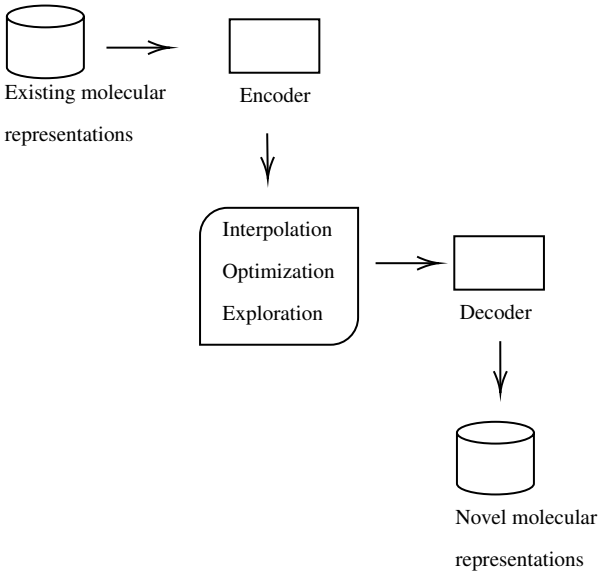

Figure 1. PGMD-VAE architecture

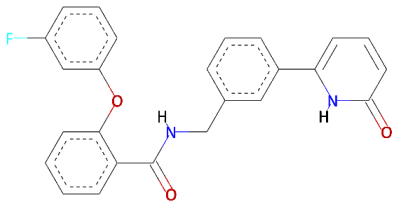

Figure 2. M0043, one of the novel candidate ligands for the SARS-CoV-2 main protease

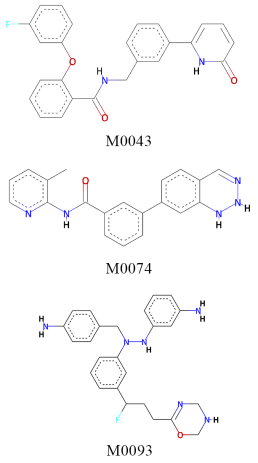

Figure 3. The three novel ligands for the SARS-CoV-2 main protease

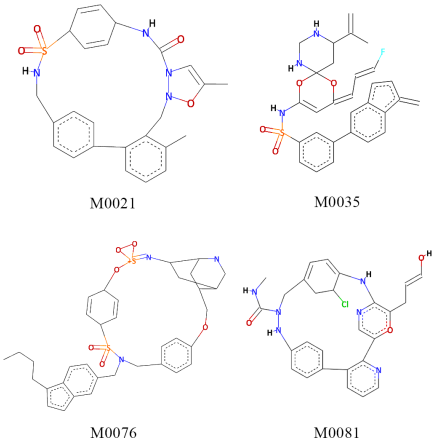

Figure 4. The four novel ligands for ACE2

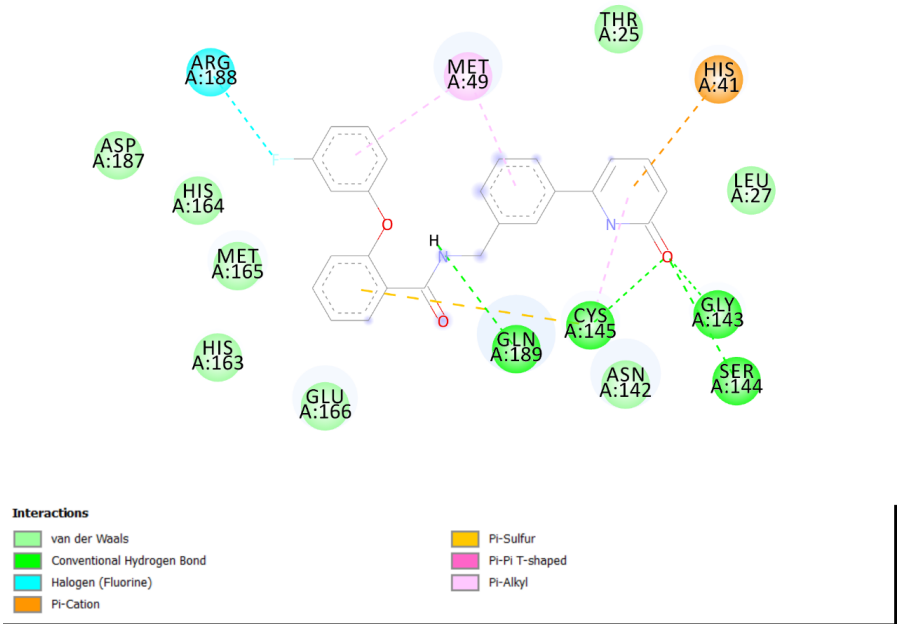

Figure 5. Interaction diagram of M0043

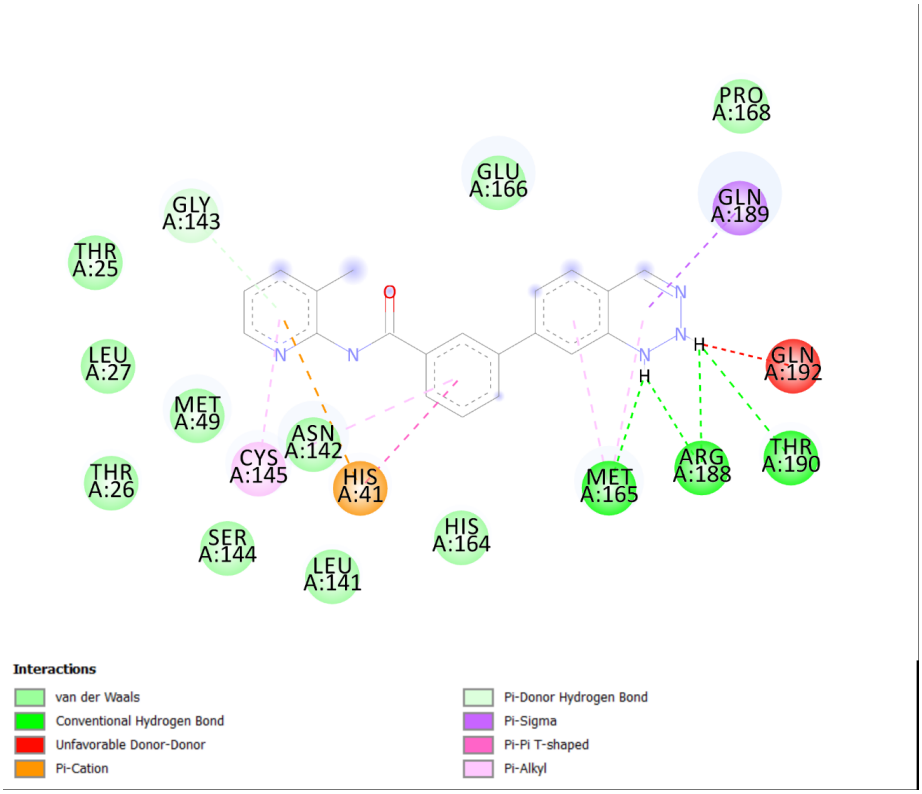

Figure 6. Interaction diagram of M0074

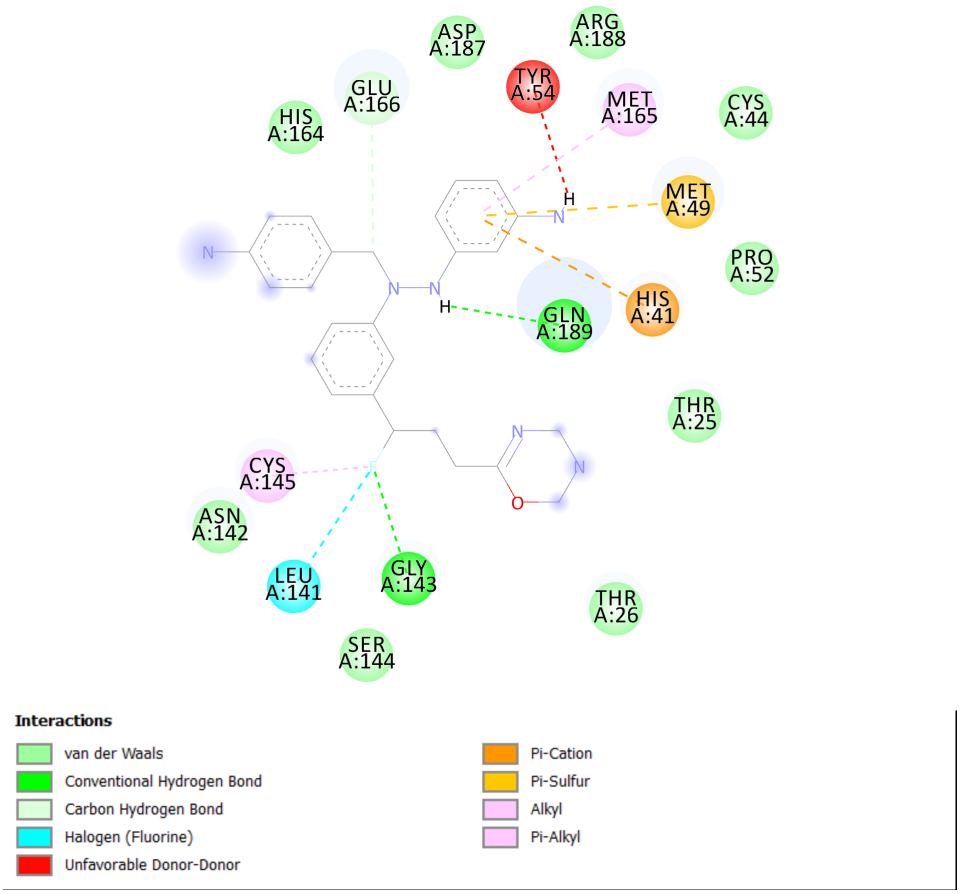

Figure 7. Interaction diagram of M0093

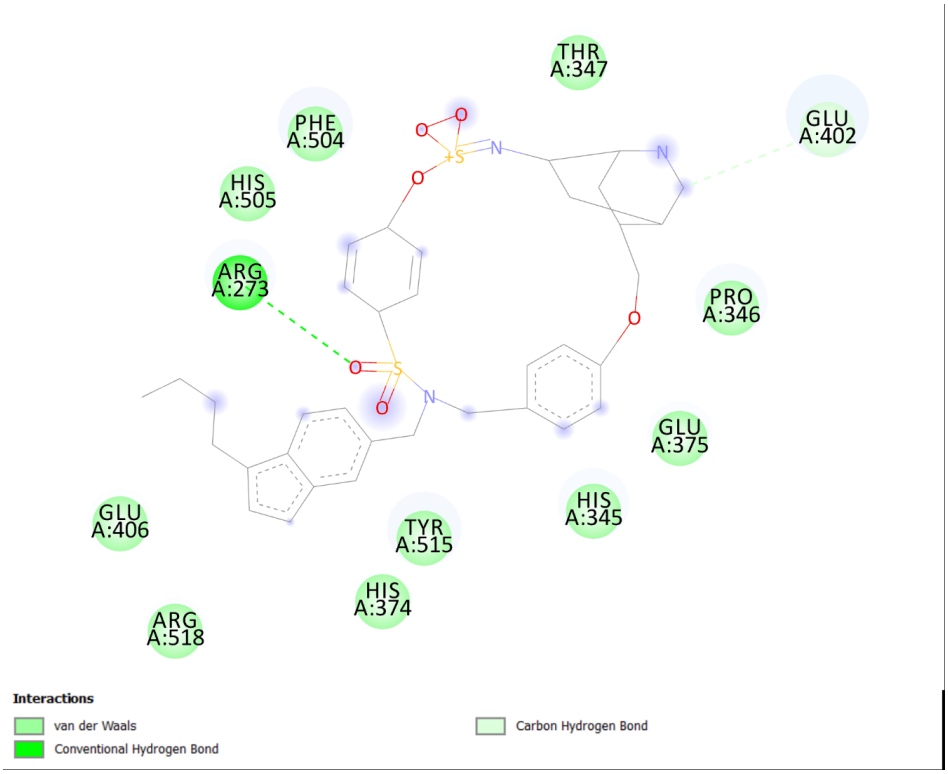

Figure 8. Interaction diagram of M0076

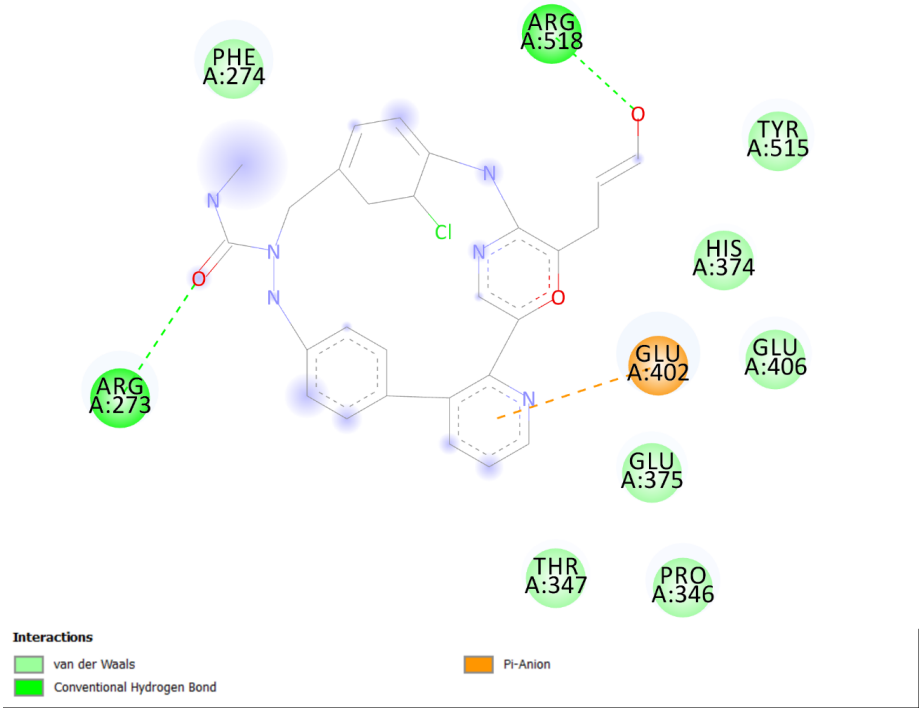

Figure 9. Interaction diagram of M0081

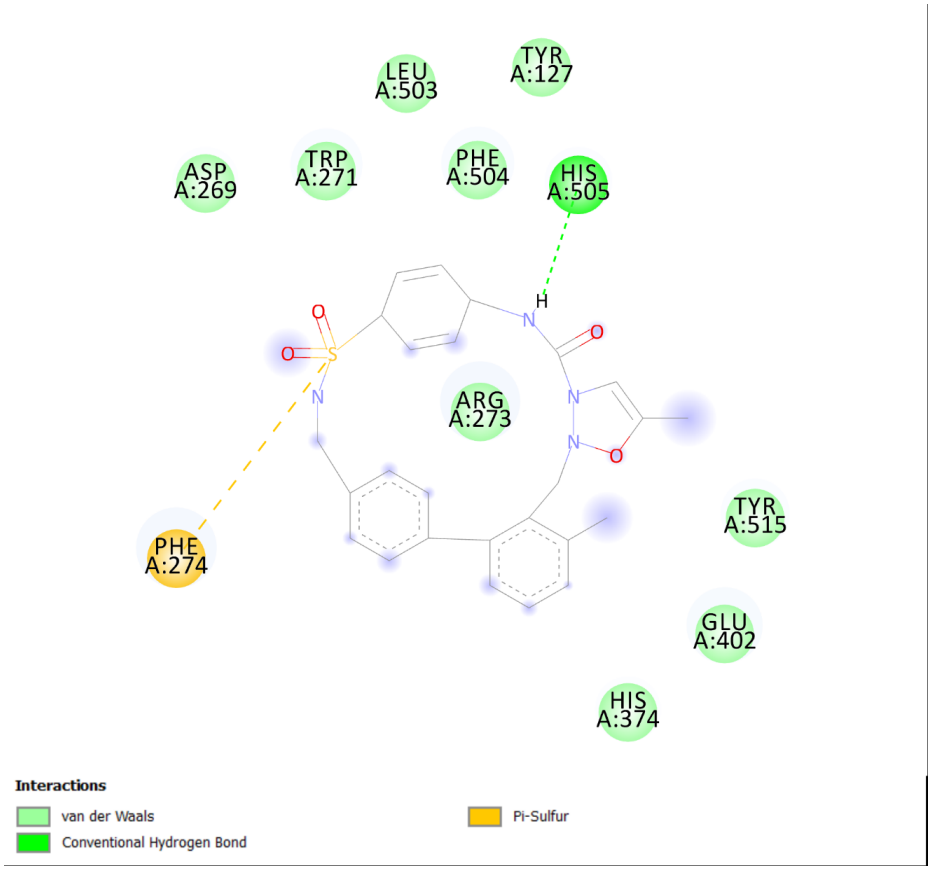

Figure 10. Interaction diagram of M0021

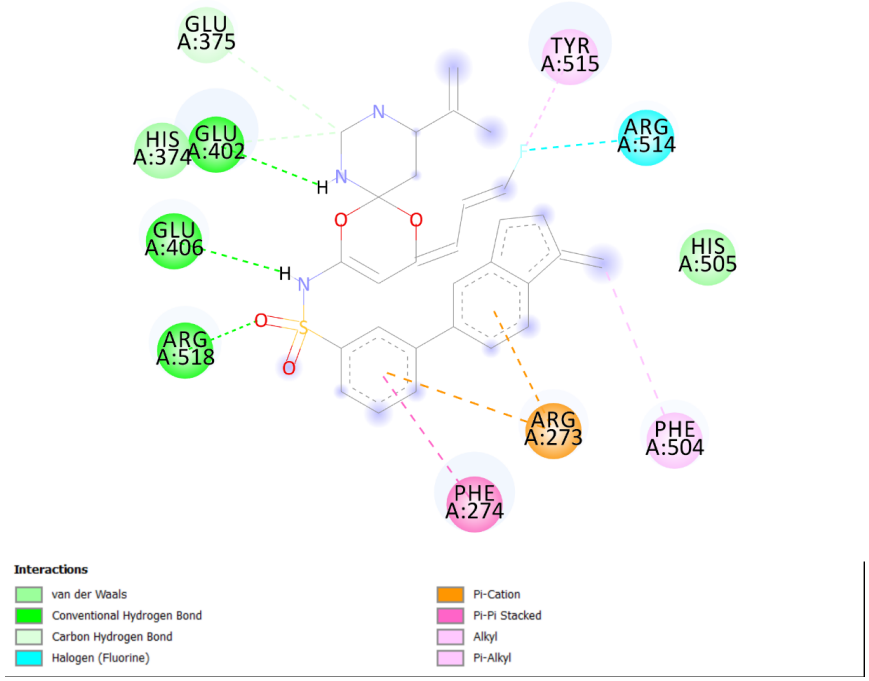

Figure 11. Interaction diagram of M0035