

# ***De novo* design of new chemical entities (NCEs) for SARS-CoV-2 using artificial intelligence**

Navneet Bung\*, Sowmya Ramaswamy Krishnan\*, Gopalakrishnan Bulusu<sup>#</sup> and Arijit Roy<sup>#</sup>  
TCS Innovation Labs-Hyderabad (Life Sciences Division), Tata Consultancy Services  
Limited, Hyderabad 500081, India.

\* Authors with equal contribution

<sup>#</sup>Authors for Correspondence: [g.bulusu@tcs.com](mailto:g.bulusu@tcs.com) and [roy.arijit3@tcs.com](mailto:roy.arijit3@tcs.com)

## **Abstract**

The novel SARS-CoV-2 is the source of a global pandemic *COVID-19*, which has severely affected the health and economy of several countries. Multiple studies are in progress, employing diverse approaches to design novel therapeutics against the potential target proteins in SARS-CoV-2. One of the well-studied protein targets for coronaviruses is the chymotrypsin-like (3CL) protease, responsible for post-translational modifications of viral polyproteins essential for its survival and replication in the host. There are ongoing attempts to repurpose the existing viral protease inhibitors against 3CL protease of SARS-CoV-2. Recent studies have proven the efficiency of artificial intelligence techniques in learning the known chemical space and generating novel small molecules. In this study, we employed deep neural network-based generative and predictive models for *de novo* design of new small molecules capable of inhibiting the 3CL protease. The generated small molecules were filtered and screened against the binding site of the 3CL protease structure of SARS-CoV-2. Based on the screening results and further analysis, we have identified 31 potential compounds as ideal candidates for further synthesis and testing against SARS-CoV-2.

**Keywords:** *COVID-19, SARS-CoV-2, 3CL protease, Artificial intelligence, Deep learning, Protease inhibitors*

## **Introduction**

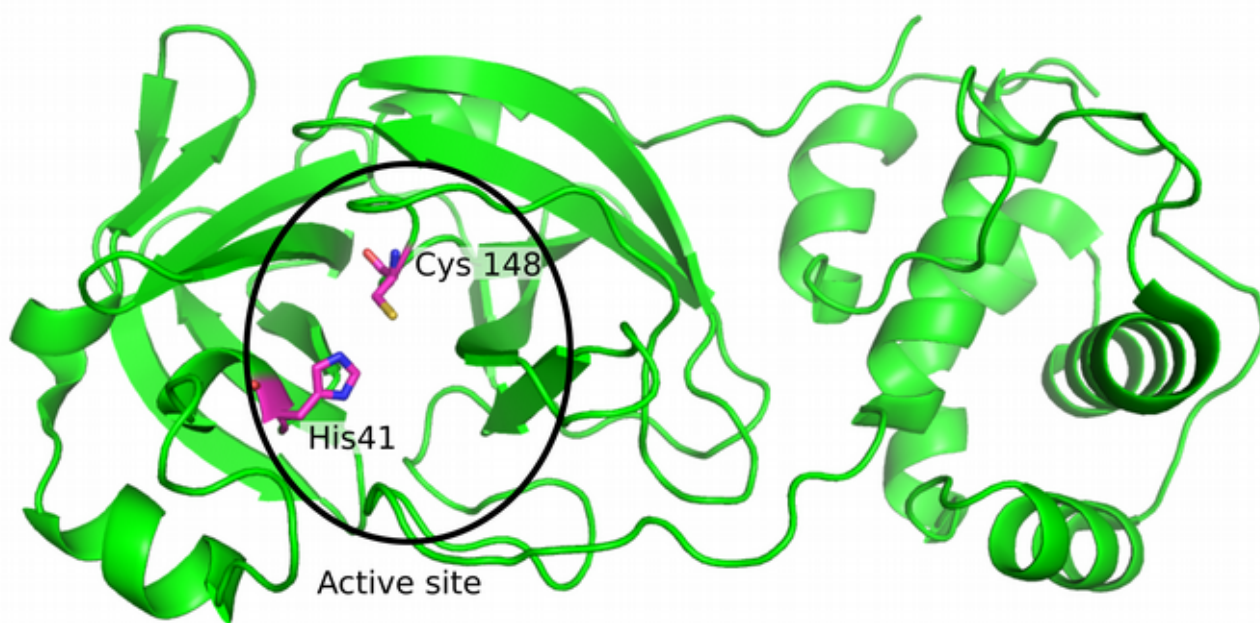
The emergence of coronaviruses has posed a severe threat to the human population. The novel coronavirus, SARS-CoV-2, has led to a worldwide pandemic (Zhou et al., 2020) with more than 168,000 cases and 6,610 deaths reported so far (WHO, 2020). In the past, there have been occurrences of epidemics involving the severe acute respiratory syndrome virus (SARS-CoV) in 2003 (Hilgenfeld and Peiris, 2013) and the Middle-East respiratory syndrome virus (MERS-CoV) in 2012 (Zaki et al., 2012). However, the absence of

promising vaccines and therapeutic drugs for these coronaviruses has decelerated the countless efforts to curb the spread of the SARS-CoV-2 worldwide.

Coronaviruses belong to the family Coronaviridae, which consist of four genera, viz., Alphacoronavirus, Betacoronavirus, Gammacoronavirus, and Deltacoronavirus. The SARS-CoV-2 belongs to the betacoronaviridae genus and its genome is ~96% identical to the bat coronavirus genome (Huang et al., 2020). The SARS-CoV-2 has the largest RNA genome of ~32 kb in size. The genome codes for both structural and non-structural proteins that contribute to viral replication, viral assembly and host-pathogen interactions. The major structural proteins of all coronaviruses are: spike glycoprotein (S), envelope proteins (E), membrane proteins (M) and nucleocapsid proteins (N) (Zhavoronkov et al., 2020; Wu et al., 2020). The spike glycoproteins play a crucial role in the viral entry into the host cells by interacting with the host cell surface receptors such as the angiotensin-converting enzyme 2 (ACE2) and the serine protease TMPRSS2 (Letko and Munster, 2020; Hoffmann et al., 2020). Upon infection the virus directs the synthesis of two long polyproteins that are needed for its replication. The viral proteases play an essential role in cleaving these polyproteins into various functional proteins. The chymotrypsin-like (3CL) protease, also known as M<sup>pro</sup> or main protease, is primarily responsible for cleaving the polyproteins, while the papain-like protease also aids in the process (Fan et al., 2004). The spike protein, 3CL protease and the papain-like protease are few of the promising drug targets to curb SARS-CoV-2 (Wu et al., 2020). Apart from the viral proteins, host proteins that interact with the virus including ACE2 and TMPRSS2 also serve as promising drug targets (Hoffmann et al., 2020).

There are ongoing efforts to synthesize small proteins that can inhibit the spike protein of SARS-Cov-2 (Begley, 2020). Nevertheless, viral proteases have been the attractive target for ongoing therapeutic development efforts. Several researchers and pharmaceutical companies are attempting to repurpose existing anti-viral drugs for the novel SARS-CoV-2 (Harrison, 2020). The protease inhibitors such as darunavir, lopinavir, ritonavir, indinavir, saquinavir and ASC09 are in clinical trials against SARS-CoV-2 (Harrison, 2020). With the recent developments in the field of artificial intelligence (AI), it is possible to mine existing knowledge and use this information to explore the virtually unlimited chemical space and develop novel small molecules with desired biological and physicochemical properties (Popova et al., 2018; Zhavoronkov et al., 2019; Stokes et al., 2020). Recently, the AI-based methods have been used to develop novel anti-bacterial molecules (Stokes et al. 2020).

In this study, we have utilized our in-house deep neural network-based generative and predictive models to design novel drug-like small molecules (new chemical entities or NCEs) against the 3CL protease of the SARS-CoV-2. The 3CL protease is a homodimeric cysteine protease (Liu et al., 2020). The crystal structure of 3CL protease is available in protein data bank (PDB: 6LU7, Fig. 1) (Liu et al., 2020). The 3CL protease of SARS-CoV-2 is 96% identical to the 3CLP protease from SARS-CoV indicating the conserved nature of the protein structure. Inhibiting the 3CL protease would help to restrict the viral maturation thereby reducing the SARS-CoV-2 infection in humans.



**Figure 1:** Structure of 3CL protease from SARS-CoV-2. The active site residues, His41 and Cys148 are shown in magenta sticks.

## Methods

The deep neural network-based generative and predictive models were used to design NCEs to target the 3CL protease of the SARS-CoV-2 (Fig. 2).

### **Data preprocessing**

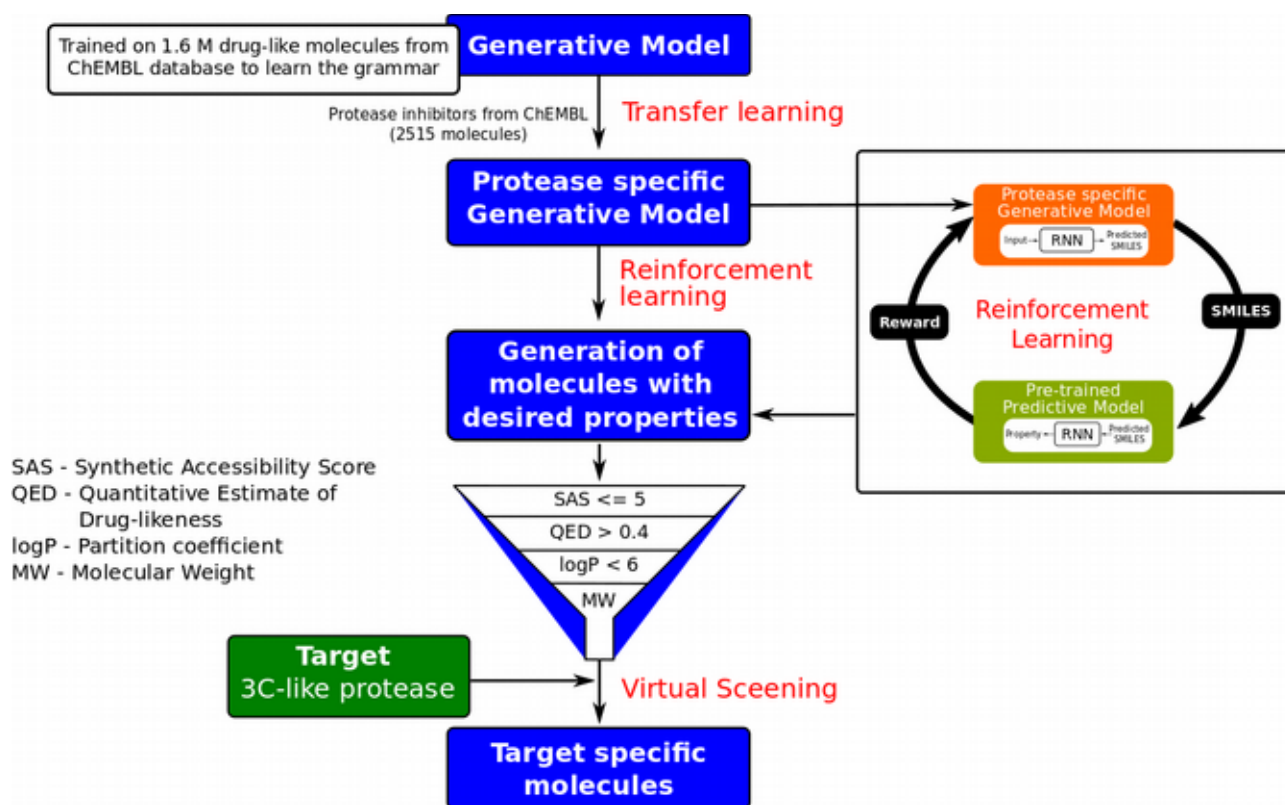
A dataset of ~1.6 million drug-like small molecules from the ChEMBL database (Gaulton et al., 2012) was used for training the generative deep neural network model. The molecules were represented in Simplified Molecular Input Line Entry System (SMILES) format (Weininger et al., 1989). The SMILES dataset was preprocessed by applying sequential filters to remove stereochemistry, salts, undesirable atoms or groups and to maintain a canonical representation throughout the training and validation process. The SMILES

strings that were greater than 100 symbols in length were also removed (Popova et al., 2018). The RDKit library in python was used for the dataset preprocessing.

### ***De novo design pipeline***

The generative model was initially trained on a dataset of ~1.6 million drug-like molecules from ChEMBL database (Gaulton et al., 2012). The generative model takes as input a dataset of small molecules in SMILES format for training. The problem of learning the SMILES grammar and reproducing it to generate novel small molecules was cast as a classification problem. The entire SMILES string was considered as a time series, where every position or symbol was considered as a time point. The different symbols in the SMILES vocabulary were considered as the classes of the classification. At a given time point, the generative model was trained to predict the class of the next symbol given the class distributions of the previous symbols in the time series. Thus, the model learns the probability distribution over the various classes at each time point of the time series.

A dataset of protease inhibitor molecules was extracted from ChEMBL database (Gaulton et al., 2012). While constructing the dataset, the molecules with a pChEMBL score greater than 7 (calculated from  $IC_{50}$ ) were considered. The dataset of 2515 protease inhibitor molecules was used to re-train the generative model using transfer learning (Tan et al., 2018). In the process of transfer learning, the probability distribution is learned by only the final few layers in order to bias the generative model towards focusing on a smaller subset of the chemical space. Further, reinforcement learning was used to modulate the generative model to produce molecules with desired properties (Sutton and Barton, 1998; Jaques et al., 2016; Olivecrona et al., 2017). The AI-generated small molecules were filtered based on various physicochemical properties. Finally, the filtered small molecules were docked to the energy minimized 3CL protease structure (PDB ID: 6LU7) and ranked based on their virtual screening scores obtained using AutoDock Vina (Trott and Olson, 2010).



**Figure 2:** AI based *de novo* NCE design pipeline of TCS.

## Results

The trained generative model was used for sampling 50,000 small molecules from the learned chemical space. After removal of duplicates and molecules which were identical to the ChEMBL database, the residual dataset consisted of 42,484 molecules. These molecules were subjected to stringent physicochemical property filters including drug-likeness (Bickerton et al., 2012) and synthetic accessibility (Ertl and Schuffenhauer, 2009), which resulted in a set of 3,960 molecules. These molecules were considered as potential candidates for inhibition of SARS-CoV-2 3CL protease. The generated molecules were screened for their affinity towards the protease using AutoDock Vina (Trott and Olson, 2010). After virtual screening, a total of 1,333 small molecules were obtained that had a virtual screening score below -7.0 (see supplementary information).

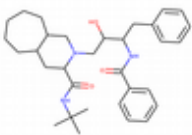
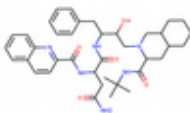
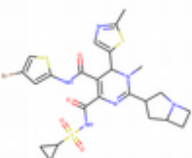
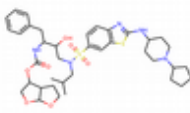
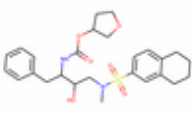
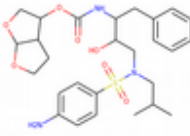
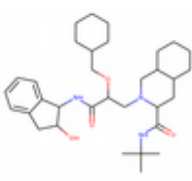
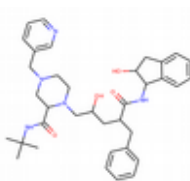
### ***Generated NCEs have similarity to protease inhibitors currently in clinical trials and show better virtual screening scores***

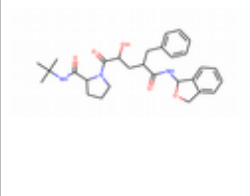
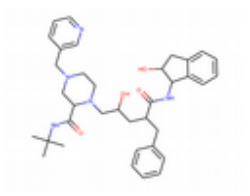
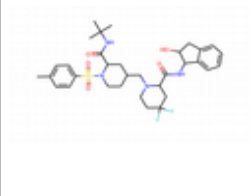
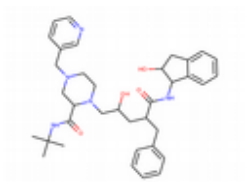
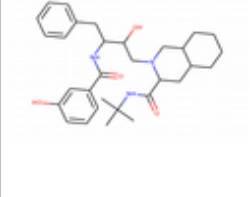
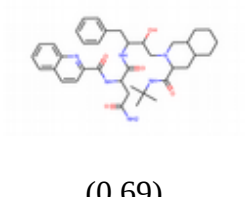
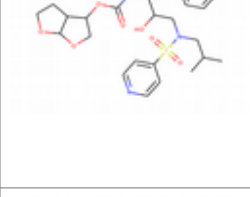
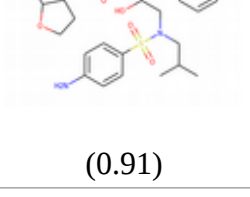
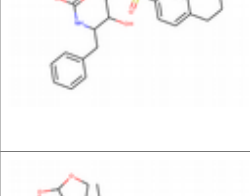
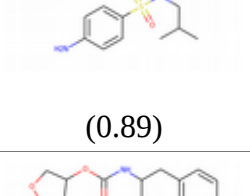
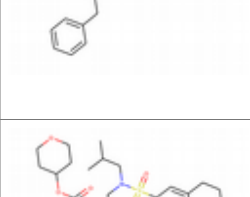
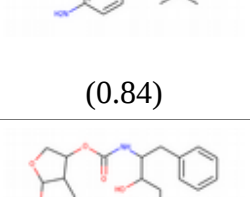
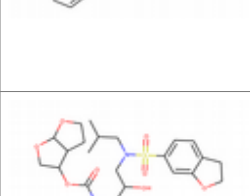
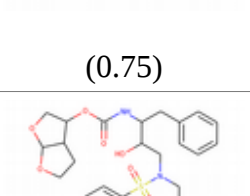


The virtual screening scores of protease inhibitors (Indinavir, Saquinavir, Darunavir, ASC09, Ritonavir and Lopinavir) currently in clinical trials for SARS-CoV-2 (Harrison,

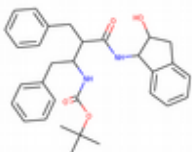
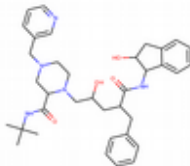
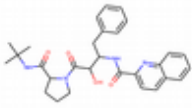
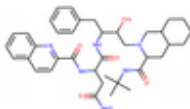
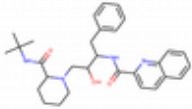
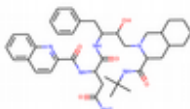
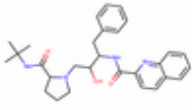
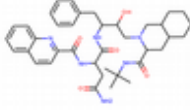
2020) were calculated using AutoDock Vina (Trott and Olson, 2010). The generated small molecules with good similarity to the molecules in clinical trials and also with a higher virtual screening score than the latter were selected for further studies (Table 1).

The highest and lowest virtual screening scores were -8.3 and -7.5 respectively. The highest tanimoto coefficient (Lipkus, 1999) was found to be 0.91 with the HIV-protease inhibitor Darunavir (TCS\_RnI\_32855). Most of the top 16 molecules were found to have higher similarity to Darunavir, Indinavir and Saquinavir.

**Table 1:** NCEs that are similar to the protease inhibitors (darunavir, lopinavir, ritonavir, indinavir, saquinavir and ASC09 that are in clinical trials against SARS-CoV-2) but with better virtual screening score.

Molecule ID	2D Structure	Virtual screening score	Protease inhibitors in trials (Tanimoto similarity)	Virtual screening score of protease inhibitors in trials
TCS_RnI_32195		-8.3	 (0.70)	-6.9
TCS_RnI_11071		-8.2	 (0.60)	-7.9
TCS_RnI_41874		-8.1	 (0.74)	-7.8
TCS_RnI_16035		-8.1	 (0.68)	-7.8

TCS_RnI_27567		-8.1	 (0.61)	-7.8
TCS_RnI_04004		-8.1	 (0.60)	-7.8
TCS_RnI_37788		-8.1	 (0.69)	-6.9
TCS_RnI_32855		-8	 (0.91)	-7.8
TCS_RnI_29372		-7.9	 (0.89)	-7.8
TCS_RnI_33357		-7.9	 (0.84)	-7.8
TCS_RnI_05476		-7.9	 (0.75)	-7.8
TCS_RnI_36110		-7.8		-7.8

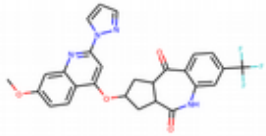
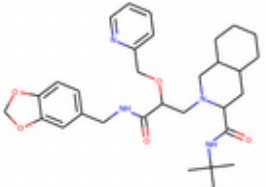
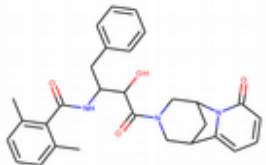
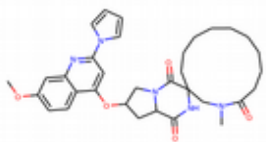
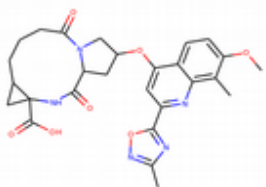
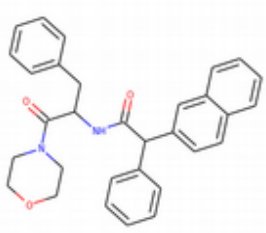
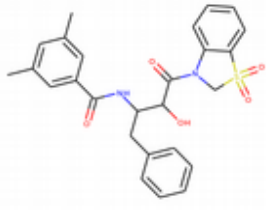
			(0.85)	
TCS_RnI_25474		-7.8	 (0.62)	-7.8
TCS_RnI_36487		-7.7	 (0.80)	-6.9
TCS_RnI_15219		-7.7	 (0.83)	-6.9
TCS_RnI_30090		-7.5	 (0.81)	-6.9

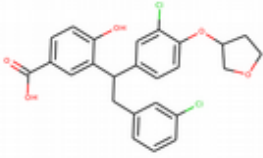
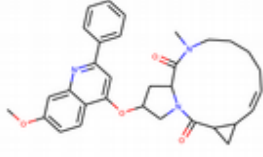
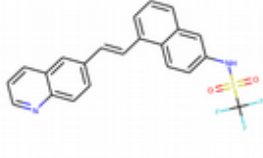
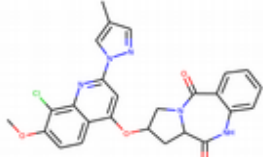
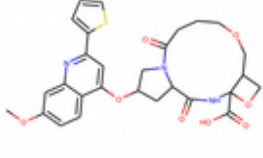
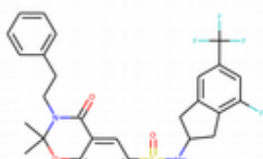
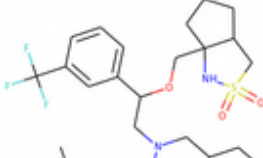
### **Generated small molecules with high virtual screening scores**

The molecules with virtual screening score below -8.5 were considered as potential candidates for synthesis and testing against SARS-CoV-2 (Table 2). The highest virtual screening score was found to be -9.1 (TCS\_RnI\_28301) while the highest tanimoto coefficient (Lipkus, 1999) with the existing protease inhibitors among the top 15 molecules was found to be 0.90 (TCS\_RnI\_40160) (Table 1). Five of the top 15 compounds had a high virtual screening score as well as higher similarity to the existing protease inhibitors (tanimoto coefficient > 0.80). The internal diversity (Benhenda, 2017) of the compounds was found to be 0.61 indicating that, the molecules generated were similar in 40% of their structural features which might contribute to their binding to proteases and, sufficiently diverse to span the spectrum of proteases identified.

**Table 2:** NCEs with highest virtual screening score against 3CLP protease of SARS-CoV-2. The tanimoto similarity to the known protease inhibitors along with their corresponding ChEMBL ID has been provided.



Molecule ID	2D Structure	Virtual screening score	ChEMBL ID* (Tanimoto similarity)
TCS_RnI_28301		-9.1	CHEMBL43030 (0.64)
TCS_RnI_21135		-8.9	CHEMBL99607 (0.58)
TCS_RnI_38094		-8.8	CHEMBL1762146 (0.65)
TCS_RnI_02739		-8.7	CHEMBL288627 (0.81)
TCS_RnI_25998		-8.7	CHEMBL1241340 (0.79)
TCS_RnI_34984		-8.7	CHEMBL3326424 (0.64)
TCS_RnI_35690		-8.7	CHEMBL1762146 (0.59)

TCS_RnI_12310		-8.7	CHEMBL48083 (0.51)
TCS_RnI_29997		-8.6	CHEMBL222212 (0.85)
TCS_RnI_22096		-8.6	CHEMBL293998 (0.84)
TCS_RnI_03318		-8.6	CHEMBL3634160 (0.77)
TCS_RnI_40647		-8.6	CHEMBL414076 (0.75)
TCS_RnI_26851		-8.6	CHEMBL1762146 (0.62)
TCS_RnI_23476		-8.6	CHEMBL2403638 (0.60)

TCS_RnI_40160		-8.4	CHEMBL3906508 (0.90)
---------------	---	------	-------------------------

**\* ChEMBL ID of known protease inhibitor that has highest similarity to the AI-generated molecule.**

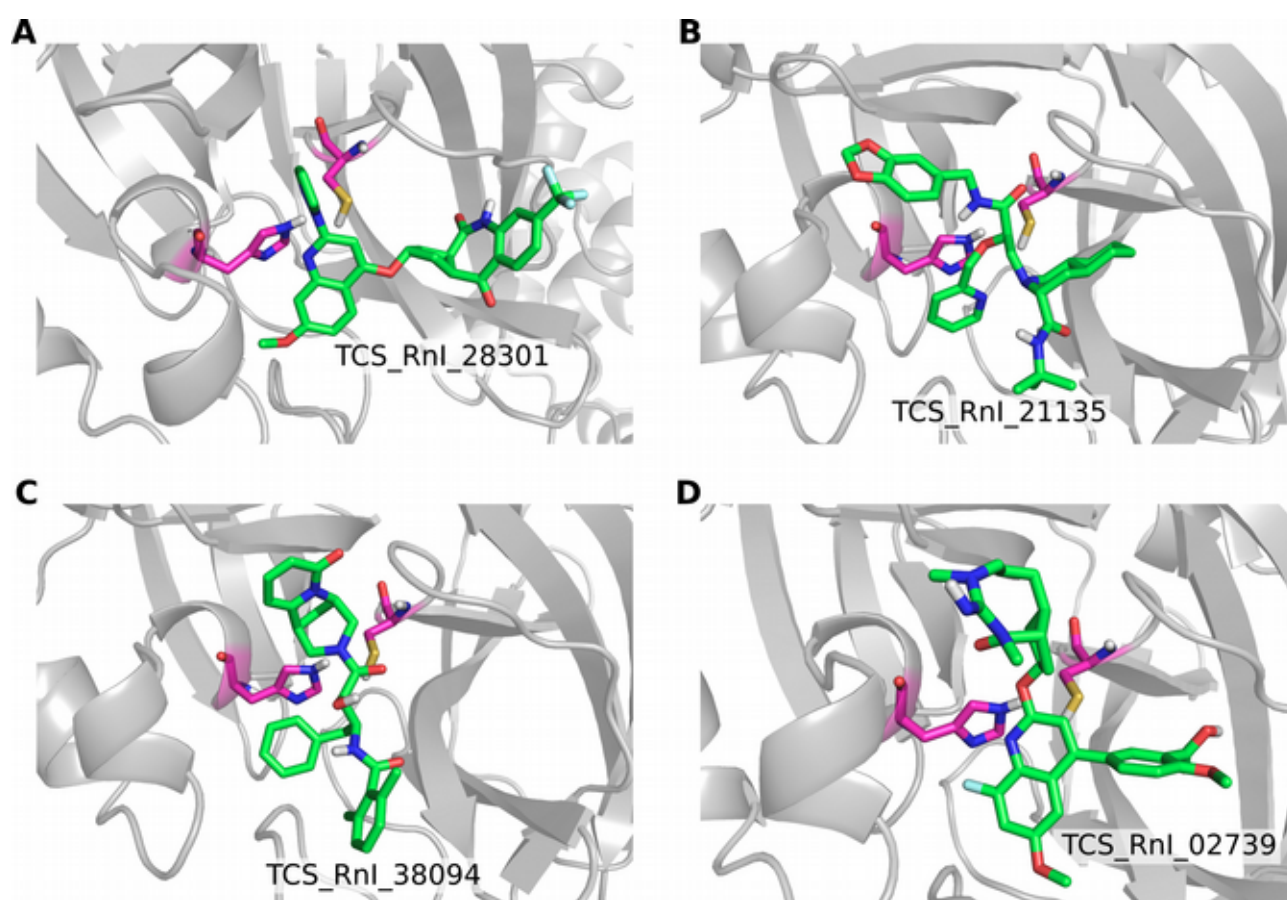
The various physicochemical properties of the top 31 small molecules (Tables 1 and 2) from our generative model are presented in Table 3.

**Table 3:** Physico-chemical properties of the top 31 compounds shortlisted for synthesis and testing (NCEs from Tables 1 and 2).

Molecule ID	logP*	MW* (Da)	HBA*	HBD*	NRB*	TPSA* (Å <sup>2</sup> )	Benzene count	QED*	SAS*
TCS_RnI_28301	5.06	522.2	7	1	4	95.3	4	0.40	3.90
TCS_RnI_21135	3.81	550.3	7	2	9	102.0	2	0.49	3.99
TCS_RnI_38094	2.74	485.2	5	2	6	91.6	3	0.56	4.41
TCS_RnI_02739	4.23	587.3	7	1	4	106.0	3	0.49	4.94
TCS_RnI_25998	2.40	535.2	9	2	5	157.0	3	0.50	4.64
TCS_RnI_34984	4.56	478.2	3	1	7	58.6	4	0.42	2.90
TCS_RnI_35690	2.78	478.2	5	2	6	103.8	3	0.57	3.35
TCS_RnI_12310	5.94	472.1	4	2	7	76.0	3	0.44	3.22
TCS_RnI_29997	5.09	525.3	5	0	4	72.0	3	0.44	4.35
TCS_RnI_22096	5.82	428.1	3	1	4	59.1	4	0.42	2.46
TCS_RnI_03318	4.01	503.1	7	1	4	98.6	4	0.45	3.64
TCS_RnI_40647	2.67	567.2	9	2	5	136.5	3	0.48	4.79
TCS_RnI_26851	5.10	562.2	4	1	6	75.7	3	0.42	3.40
TCS_RnI_23476	5.03	599.3	5	2	7	87.7	1	0.46	4.84
TCS_RnI_40160	4.15	437.1	7	1	4	100.9	5	0.46	2.97
TCS_RnI_32195	4.57	519.3	4	3	8	81.7	2	0.48	3.68
TCS_RnI_11071	3.26	638.0	10	2	7	124.1	2	0.48	4.72
TCS_RnI_41874	2.67	502.2	6	2	9	105.2	2	0.55	3.52
TCS_RnI_16035	4.52	553.4	5	3	8	90.9	1	0.44	4.31
TCS_RnI_27567	2.85	507.3	5	3	8	108.0	2	0.51	3.92
TCS_RnI_04004	3.55	646.3	6	3	7	119.1	2	0.42	4.33
TCS_RnI_37788	3.89	521.3	5	4	8	101.9	2	0.42	3.79
TCS_RnI_32855	2.19	533.2	8	2	11	127.3	2	0.45	4.02
TCS_RnI_29372	3.67	586.3	7	2	11	114.4	2	0.41	4.07
TCS_RnI_33357	3.12	588.3	8	2	11	123.6	2	0.41	4.15
TCS_RnI_05476	4.09	558.3	6	2	11	105.2	2	0.43	3.41
TCS_RnI_36110	2.73	574.2	8	2	11	123.6	2	0.42	4.13
TCS_RnI_25474	4.76	500.3	4	3	8	87.7	3	0.42	3.51
TCS_RnI_36487	2.84	502.3	5	3	7	111.6	3	0.46	3.33
TCS_RnI_15219	3.71	502.3	5	3	8	94.6	3	0.44	3.37
TCS_RnI_30090	3.32	488.3	5	3	8	94.6	3	0.45	3.36

\*logP – Partition coefficient; MW – Molecular Weight; HBA – Hydrogen Bond Acceptor; HBD – Hydrogen Bond Donor; TPSA – Topological Polar Surface Area; QED – Quantitative Estimate of Drug-likeness; SAS – Synthetic Accessibility Score.

The best docking pose for the top 4 compounds from Table 2 are shown in Figure 3. Comparative analysis with known protease inhibitors and molecules in clinical trials showed that the generative model was able to capture the features of the protease inhibitors governing binding affinity, such as favorable substituents and scaffolds. These features are expected to aid in the inhibition of SARS-CoV-2 3CL protease resulting in the reduction of the viral infection in human body.



**Figure 3:** The best docking pose of the top 4 compounds generated using AutoDock Vina (A-D). The SARS-CoV-2 3CL protease is shown in cartoon representation while the top four compounds are shown as green sticks. The active site residues, His41 and Cys145, are shown as magenta sticks.

## Summary

In this work, we have designed small molecules that can inhibit the SARS-CoV-2 3CL protease, which is responsible for viral replication. We have harnessed the power of deep learning methods to learn the inherent grammar of small molecules and produce novel molecules, which satisfy the learnt grammar. We have also used various physico-chemical property filters such that, the generated molecules possess drug-like properties. Finally, virtual screening was performed to obtain a ranked list of molecules. We also observed that, the generative model could generate small molecules that are similar to HIV-protease inhibitors, but better binding to SARS-CoV-2 3CL protease (Table 1). A list of small molecules, which show good virtual screening score has also been provided (Table 2). The complete set of promising small molecules can be found in the supporting information, so that anyone can test these molecules against SARS-CoV-2 in this hour of need.

One of the main challenges for us is to synthesize these molecules and test them against SARS-CoV-2. We are collaborating with our academic partners, particularly the laboratories of the Council of Scientific and Industrial Research (CSIR), India for the same.

## Acknowledgments

We thank our colleagues Dr. Rajgopal Srinivasan, Dr. Gautam Shroff, Dr. Sanjay Bhat, Dr. Siladitya Padhi, Dibyajyoti Das, Dr. Broto Chakrabarty and Sumedh Gupte for helpful discussions. We also thank Dr. Shekhar Mande, CSIR-India for his interest in this work.

## Supplementary Material

A list of all promising molecules along with their virtual screening scores against 3CL protease of SARS-CoV-2.

## References

- Begley S. <https://www.statnews.com/2020/03/09/coronavirus-scientists-play-legos-with-proteins-to-build-next-gen-vaccine/>. 2020.
- Benhenda M. ChemGAN challenge for drug discovery: can AI reproduce natural chemical diversity?. *ArXiv*. 2017.
- Bickerton G R, Paolini G V, Besnard J, Muresan S, and Hopkins A L. Quantifying the chemical beauty of drugs. *Nat Chem*. 2012. 4(2). 90-98.
- Ertl P and Schuffenhauer A. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *J Cheminform*. 2009. 1(1). 8.

Fan K, Wei P, Feng Q, Chen S, Huang C, Ma L et al. Biosynthesis, Purification, and Substrate Specificity of Severe Acute Respiratory Syndrome Coronavirus 3C-like Proteinase. *J Biol Chem*. 2004. 279(3). 1637-1642.

Gaulton A, Bellis L J, Bento A P, Chambers J, Davies M, Hersey A et al. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res*. 2012. 40(D1). D1100-D1107

Harrison C. Coronavirus puts drug repurposing on the fast track. *Nature*. 2020.

Hilgenfeld R and Peiris M. From SARS to MERS: 10 years of research on highly pathogenic human coronaviruses. *Antiviral Res*. 2013. 100(1). 286–295.

Hoffmann M, Kleine-Weber H, Schroeder S, Krüger N, Herrler T, Erichsen S et al. SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor. *Cell*. 2020. doi: 10.1016/j.cell.2020.02.052.

Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet*. 2020. doi: 10.1016/S0140-6736(20)30183-5.

Jaques N, Gu S, Turner R E and Eck D. Tuning recurrent neural networks with reinforcement learning. *ArXiv*. 2016. 1609.05473

Lipkus A H. A proof of the triangle inequality for the Tanimoto distance. *Journal of Mathematical Chemistry*. 1999. 26(1-3). 263-265

Liu X, Zhang B, Jin Z, Yang H and Rao Z. The crystal structure of COVID-19 main protease in complex with an inhibitor N3. 2020. *To be published*.

Olivecrona M, Blaschke T, Engkvist O and Chen H. Molecular *de-novo* design through deep reinforcement learning. *J Cheminform*. 2017. 9(1). 48

Popova M, Isayev O and Tropsha A. Deep reinforcement learning for de novo drug design. *Sci. Adv*. 2018. 4(7). eaap7885.

RDKit: Open-source cheminformatics; <http://www.rdkit.org>

Stokes J M, Yang K, Swanson K, Jin W, Cubillos-Ruiz A, Donghia NM, et al. A Deep Learning Approach to Antibiotic Discovery. *Cell*. 2020. 180. 688-702.

Sutton R and Barton A. Reinforcement learning: an introduction, 1<sup>st</sup> edn. *MIT Press, Cambridge*. 1998

Tan C, Sun F, Kong T, Zhang W, Yang C and Liu C. A survey on deep transfer learning. *ArXiv*. 2018.

Trott O and Olson A J. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization and multithreading. *J Comput Chem*. 2010. 31(2). 455-461

Weininger D, Weininger A and Weininger J L. Smiles. 2. Algorithm for generation of unique smiles notation. *J. Chem. Inf. Comput. Sci*. 1989. 29(2). 97-101.

WHO. <https://experience.arcgis.com/experience/685d0ace521648f8a5beeeee1b9125cd>, accessed on March 17, 2020.

Wu C, Liu Y, Yang Y, Zhang P, Zhong W, Wang Y et al. Analysis of therapeutic targets for SARS-CoV-2 and discovery of potential drugs by computational methods. *Acta Pharmaceutica Sinica B*. 2020. doi: 10.1016/j.apsb.2020.02.008.

Zaki A M, van Boheemen S, Bestebroer T M, Osterhaus A D and Fouchier R A. Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia. *N. Engl. J. Med*. 2012(19). 367. 1814–1820.

Zhavoronkov A, Aladinskiy V, Zhebrak A, Zagribelnyy B, Terentiev V, Bezrukov DS et al. Potential 2019-nCoV 3C-like protease inhibitors designed using generative deep learning approaches. *Chemrxiv* .2020. doi: 10.26434/chemrxiv.11829102.v2.

Zhavoronkov A, Ivanenkov Y A, Aliper A, Veselov M S, Aladinskiy V A, Aladinskaya A V, et al. Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nat. biotechnol*. 2019. 37(9). 1038-40.

Zhou P, Yang X L, Wang X G, Hu B, Zhang L, Zhang W et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature*. 2020. doi: 10.1038/s41586-020-2012-7.