NOTE: None of the molecules have been synthesized or tested in vitro or in vivo. These are not drugs for 2019-nCov coronavirus. Expert medicinal chemists are encouraged to review and comment on the molecules in the article and on the website.

Potential 2019-nCoV 3C-like protease inhibitors designed using generative deep learning approaches

Alex Zhavoronkov, Vladimir Aladinskiy, Alexander Zhebrak, Bogdan Zagribelnyy, Victor Terentiev, Dmitry S. Bezrukov, Daniil Polykovskiy, Rim Shayakhmetov, Andrey Filimonov, Philipp Orekhov, Yilin Yan, Olga Popova, Quentin Vanhaelen, Alex Aliper, Yan Ivanenkov

Insilico Medicine Hong Kong Ltd, Pak Shek Kok, New Territories, Hong Kong

Corresponding author: Alex Zhavoronkov, email: alex@insilico.com

The emergence of the 2019 novel coronavirus (2019-nCoV), for which there is no vaccine or any known effective treatment created a sense of urgency for novel drug discovery approaches. One of the most important 2019-nCoV protein targets is the 3C-like protease for which the crystal structure is known. Most of the immediate efforts are focused on drug repurposing of known clinically-approved drugs and virtual screening for the molecules available from chemical libraries that may not work well. For example, the IC50 of lopinavir, an HIV protease inhibitor, against the 3C-like protease is approximately 50 micromolar. In an attempt to address this challenge, on January 28th, 2020 Insilico Medicine decided to utilize a part of its generative chemistry pipeline to design novel drug-like inhibitors of 2019-nCoV and started generation on January 30th. It utilized three of its previously validated generative chemistry approaches: crystal-derived pocked-based generator, homology modelling-based generation, and ligand-based generation. Novel druglike compounds generated using these approaches are being published at www.insilico.com/ncov-sprint/ and will be continuously updated. Several molecules will be synthesized and tested using the internal resources; however, the team is seeking collaborations to synthesize, test, and, if needed, optimize the published molecules.

Introduction

Coronaviruses (CoVs) are a large family of viruses belonging to the family Coronaviridae. The limited number of coronaviruses known to be circulating in humans cause mild infections and

they were regarded as relatively harmless respiratory human pathogens [1]. The emergence of the severe acute respiratory syndrome coronavirus (SARS-CoV) and the Middle East Respiratory Syndrome (MERS) virus revealed that coronaviruses can cause severe and sometimes fatal respiratory tract infections in humans. The first known case of SARS-CoV occurred in Foshan, China in November 2002 and new cases emerged in mainland China in February 2003. The first emergence of MERS-CoV occurred in June 2012 in Saudi Arabia [2]. These events demonstrated that the threats of CoVs should not be underestimated and that it is of paramount importance to advance the knowledge on the replication of these viruses and their interactions with the hosts to develop treatments and vaccines. These successive outbreaks also highlight the long-term threat of cross-species transmission events leading to outbreaks in humans and the possible re-emergence of similar virus infection that should be considered seriously [3]. SARS-CoV and MERS-CoV are two major causes of severe atypical pneumonia in humans and share important features that contribute to preferential viral replication in the lower respiratory tract and viral immunopathology. In December 2019, atypical pneumonia cases emerged in Wuhan, Hubei, China, with clinical presentations consistent with viral pneumonia. The cause was quickly identified as being a novel CoV, which was named 2019 novel coronavirus (2019-nCoV). Investigations of the epidemiological, clinical, laboratory and radiological characteristics, treatment, and outcomes of patients infected by 2019-nCoV demonstrated that the infection caused clusters of severe respiratory illness similar to SARS-CoV [4]. Early clinical investigations showed that although the 2019-nCoV can cause severe illness in some patients, it initially did not transmit readily between people. However, more recent epidemiological data suggest the new virus has undergone human host adaptation/evolution and has become more efficient in human to human transmission. Analysis of 2019-nCoV genome sequences obtained from patients during the beginning of the outbreak demonstrated that they are almost identical to each other and share 79.5% sequence identity to SARS-CoV [5]. The nCoV-2019 is 96% identical at the whole genome level to a bat coronavirus. The 2019-nCoV genomic sequence was used to perform comparative genetic and functional analysis with the human SARS virus and coronaviruses recovered from other species. Phylogenetic analysis of CoVs of different species indicated that 2019-nCoV could have originated from Chinese horseshoe bats, but the intermediate transmission vehicle has not yet been identified [6]. According to this study, 2019-nCoV belongs to a novel type of bat coronavirus owing to a high degree of variation from the human SARS virus. 2019-nCoV is the seventh member of the family of CoVs that infect humans. Like SARS-CoV, 2019-nCoV enters target cells through an endosomal pathway and also uses the same cell entry receptor, Angiotensin-converting enzyme II (ACE2) [5 7]. Detailed analysis of the interaction of receptor binding domains (RBDs) of 2019-nCoV with human ACE2 indicated that the affinity of binding to a human cell is lower than that of human SARS virus from which it was inferred that the infectivity and pathogenicity of this new virus could be lower than the human SARS virus [6]. Single-cell RNA expression profiling of ACE2 was carried out [8]. The analysis of the ACE2 RNA expression profile in the normal human lungs showed that the ACE2 virus receptor expression is concentrated in a small population of type II alveolar cells which is also expressed many other genes that positively regulate 2019-nCoV reproduction and transmission.

CoV structure and main strategies for targeting 2019-nCoV

The *Coronaviridae* family consists of four genera based on their genetic properties, including genus Alphacoronavirus, genus Betacoronavirus, genus Gammacoronavirus, and genus Deltacoronavirus. The coronavirus RNA genome (ranging from 26 to 32 kb) is the largest among all RNA viruses and the viral particle is about 125 nm in diameter [9]. CoVs have a complex genome expression strategy. In addition to a role in virus replication or virus assembly, many of the CoV proteins expressed in the infected cell contribute to the coronavirus-host interactions. This includes interactions with the host cell to create an optimal environment for CoV replication, alteration of the host gene expression and neutralization of the host's antiviral defenses. These coronavirus–host interactions are key to viral pathogenesis [10]. The genes for non-structural proteins constitute two-thirds of the CoV genome. Among the structural proteins, 4 are of special interest namely spike (S), envelope (E), membrane (M), and nucleocapsid (N). The S, E, and M proteins are contained within the viral membrane. The M and E proteins are involved in viral assembly, while the N protein is required for RNA genome assembly. The S protein, a surface-located trimeric glycoprotein of CoVs, plays a functional role in viral entry into host cells, viral infection, and pathogenesis and was considered as a major therapeutic target for treatments and vaccines against SARS-CoV and MERS-CoV. Therapeutics investigated at that time included peptides that block RBD-ACE2-binding and peptides that bind the S protein to inhibit the production of functional S1 and S2 subunits and the consequent fusion of the viral envelope with the host cell membrane [1].

Although CoVs share many similarities they also have undergone substantial genetic evolution. Identification of promising targets for antiviral therapies and vaccines against 2019-nCoV should exploit the structural similarities between SARS-CoV and 2019-nCoV and focus on proteins that are highly conserved across multiple CoVs. There is an ongoing effort to ensure that all scientific materials known about 2019-nCoV such as curated data and updated research reports are available to the scientific community. For instance, the initiative https://ghddi-ailab.github.io/Targeting2019-nCoV/ supported by the Global Health Drug Discovery Institute (GHDDI) contains experimental data of CoV related studies, homology models for 2019-nCoV targets as well as for SARS-CoV and MERS-CoV protein targets. Among the many potential targets against SARS-CoV and several other CoVs, replication-related enzymes, such as protease, are highly conserved [5]. Drugs that inhibit conserved proteases are capable of preventing replication and proliferation of the virus by interfering with the post-translational processing of essential viral polypeptides. They can also reduce the risk of mutation mediated drug-resistance. This was the case for the SARS-CoV [11], as inhibitors targeting the main protease involved in replication and proliferation were the most effective means to alleviate the epidemic. Once the target is identified, computational drug repurposing procedures were launched to identify suitable drugs. Following this approach, Lopinavir and Ritonavir, two HIV-1 protease inhibitors, were identified to be capable of inhibiting SARS-CoV main protease [12]. The SARS-CoV main protease has 96.1% of similarity with the 2019-nCoV main protease, hence it can be used as a homologous target for screening drugs that inhibit the replication and proliferation of 2019-nCoV.

In this work the selected target is the C30 Endopeptidase, also referred to as the 3C-like proteinase or coronavirus 3C-like protease (3CLP) or coronavirus main protease (M^{pro}). 3CLP is a homodimeric cysteine protease and a member of a family of enzymes found in the Coronavirus polyprotein [13]. It cleaves the polyproteins into individual polypeptides that are required for replication and transcription [14] [15]. Following the translation of the messenger RNA to yield the polyproteins, the 3CLP is first auto-cleaved from the polyproteins to become a mature enzyme [16]. The 3CLP then cleaves all the 11 remaining downstream non-structural proteins. 3CLP plays a central role in the viral replication cycle and is an attractive target against the human SARS virus [17]

Computational Approaches for 2019-nCoV

Computational drug repurposing is an effective approach to find new indications for already known drugs [18] [19]. A computational drug repurposing approach typically relies on an integrated pipeline which includes a virtual screening of drug libraries to find suitable drug-target pairs using methods such as molecular similarity while homology modelling is used to model the target. Molecular docking and binding free energy calculations are used to predict drug-target interactions and binding affinity [20]. The emergence of resistance to existing antiviral drugs and re-emerging viral infections are the biggest challenges in antiviral drug discovery. The drug repurposing approach allows finding new antiviral agents within a short period to overcome the challenges in antiviral therapy. Computational drug repurposing has been used to identify drug candidates for viral infectious diseases like Ebola, ZIKA, dengue and influenza infections [21]. These methods were also used to identify potential drugs against SARS-CoV and MERS-CoV [22]; [23] and following the 2019-nCoV outbreak, computational repurposing has been applied for 2019-nCoV. The results of some of those investigations have already been reported. For instance, by looking for drugs with high binding capacity with SARS-CoV main protease, 4 small molecule drugs, Prulifloxacin, Bictegravir, Nelfinavir, and Tegobuvi, were identified as repurposing candidates against 2019-nCoV[24]. These 4 molecules were selected by high-throughput computational screening of a library of 8,000 experimental and approved drugs and small molecules obtained from Drugbank and using the structures and sequences of SARS-CoV main protease downloaded from the PDB database. Molecular similarity search was performed by using a strategy based on the similar sequences of the structure-revealed molecules. The crystal structure of the main protease monomer was used as a target protein for molecular docking and a protein-ligand interaction analysis was performed on the resulting 690 candidates. Toxins, neurologic drugs, and antitumor drugs with strong side effects were discarded from the initial set of 690 candidates leaving 50 molecules with the capability to bind the SARS-CoV main protease. After filtering for approved drugs and performing further kinetic and biochemical analysis, the four remaining drugs were Prulifloxacin, Bictegravir, Nelfinavir, and Tegobuvi. Interestingly, Nelfinavir, an HIV-1 protease inhibitor to treat HIV, was also predicted to be a potential inhibitor of 2019-nCoV main protease by another computational-based study combining homology modelling, molecular docking and binding free energy calculation [25]. In this work, the main 2019-nCoV protease structures were modeled using

the SARS homologue (PDB ID: 2GTB) as a template. Molecular docking was performed and 1903 approved drugs were tested against the model. Based on the docking score and after further three-dimensional similarity analysis, 15 drugs were selected. 10 additional new models of the main 2019-nCoV protease were used for additional docking analysis of these 15 drugs. 6 drugs (Nelfinavir, Praziquantel, Pitavastatin, Perampanel, Eszopiclone, and Zopiclone) had good binding modes and were selected for further analysis. Binding free energy calculation was performed for 4 of the 6 drugs and Nelfinavir was selected as the most promising candidate. In another recent study[26], the main 2019-nCoV protease was also used as a target to find repurposing candidates through computational screening among clinically approved medicines. The study identified a list of 10 commercial medicines that may form hydrogen bonds to key residues within the binding pocket of 2019-nCoV main protease and may also have a higher tolerance to resistance mutations.

Generative Chemistry Approaches

Considering the virtually unlimited number of chemical structures that can be generated *de novo*, conventional computational drug design approaches tend to include limited numbers of fragments and/or employ sophisticated search strategies to sample hit compounds from a predefined area of the chemical space. To enable scientists to exploit the whole drug-like chemical space, a new type of computational methods for drug discovery has been developed using the recent advances in deep learning (DL) and artificial intelligence (AI). Such techniques can automatically extract high-dimensional abstract information without the need for manual feature design and learn nonlinear mappings between molecular structures and their biological and pharmacological properties. Deep generative models can utilize large datasets for training and perform *in silico* design of *de novo* molecular structures with predefined properties [27]. The first model of this type, a molecular generator using an adversarial auto-encoder (AAE) to generate molecular fingerprints, was released in early 2017 [28]. Since then, many architectures were proposed to generate not just valid chemical structures, but also molecules matching certain bioactivity and novelty profiles as well as other features of interest. Several milestones were recently accomplished with the use of generative chemistry in drug discovery, demonstrating that it is possible to generate molecules that can be synthesized, are active *in vitro*, metabolically stable, and elicit *in vivo* activity in disease-relevant models. The first example of an *in vitro* active molecule obtained through generative chemistry was the JAK3 inhibitor [29]. Another generative model, Generative Tensorial Reinforcement Learning (GENTRL), generated discoidin domain receptor DDR1 and DDR2 inhibitors. DDR1 and DDR2 inhibitors with different property and selectivity profiles were assayed *in vitro*, followed by *in vivo* mouse experiments that validate the pharmacokinetics of DDR1 inhibitors [30]. This experiment demonstrated that generative chemistry is capable of finding novel molecular structures with optimized properties which could not be found using repurposing approaches and other standard computational methods. With a timeframe of fewer than 25 days between the initial target selection and the generation of the lead compounds, it demonstrates that this method is also time effective.

Insilico Medicine 2019-nCoV Sprint Timeline and Methods

Insilico Medicine's drug discovery system consists of three main pipelines: target discovery, small molecule drug discovery, and predictors of clinical trial outcomes (Figure 1). This system is designed to achieve maximum automation of drug discovery processes for a broad range of human diseases. Our small molecule drug discovery pipeline can be used to generate inhibitors of bacterial and viral protein targets. Multiple publications explaining the basic concepts and approaches in generative chemistry were published by the team [28–36].

Since there is a known protease target for 2019-nCoV and its sequence and structure are known, we decided to apply only the generative chemistry pipeline to generate the possible drug-like hits.

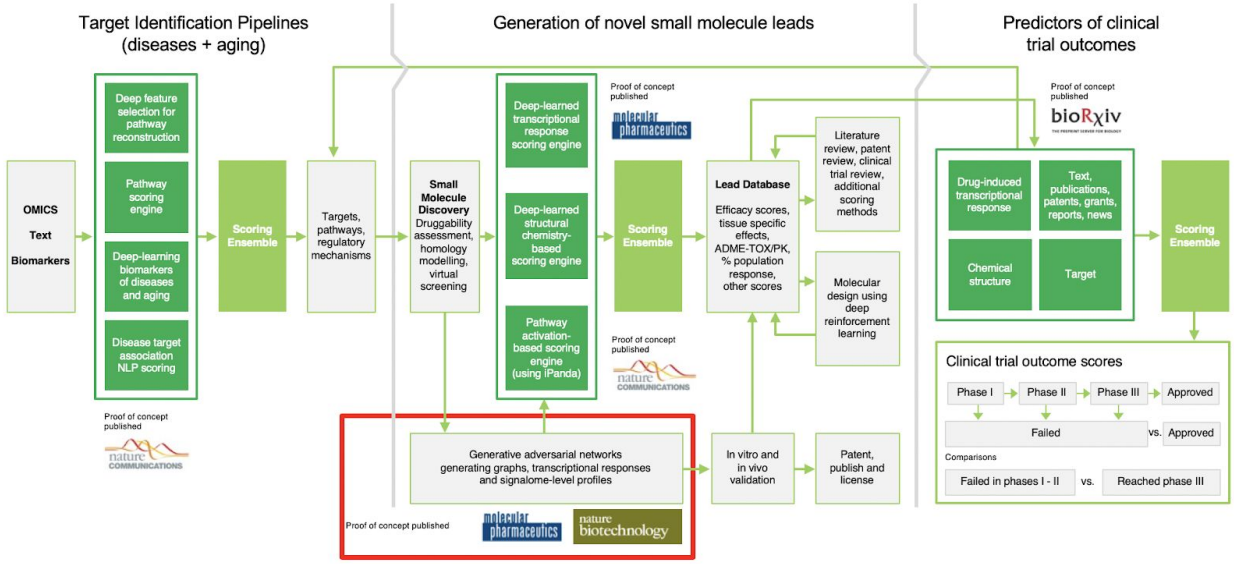

Figure 1: Insilico Medicine drug discovery pipeline. The generative modules utilizing crystal structure, homology modelling, and ligand-based generative chemistry pipelines were used to generate the molecules for the 3C-like protease.

At the end of January, the news of the 2019-nCoV showed that the virus is substantially more dangerous than previously thought. While multiple teams already proposed the most likely repurposing candidates, we decided to support the ongoing efforts with a different strategy and employed the generative chemistry approach to design novel small molecules designed specifically against 2019-nCoV. Using the 2019-nCoV 3C-like protease as a target, we planned out the generative chemistry timeline (Figure 2) starting with target selection on January 28th and publication of the molecules from the three generative approaches on February 5th. We also agreed with the key synthetic chemistry partner to start synthesizing and testing several generated molecules right after publication. Three parallel approaches were utilized to generate novel structures (pocket-based, ligand-based and homology model-based generation, represented in Figure 3).

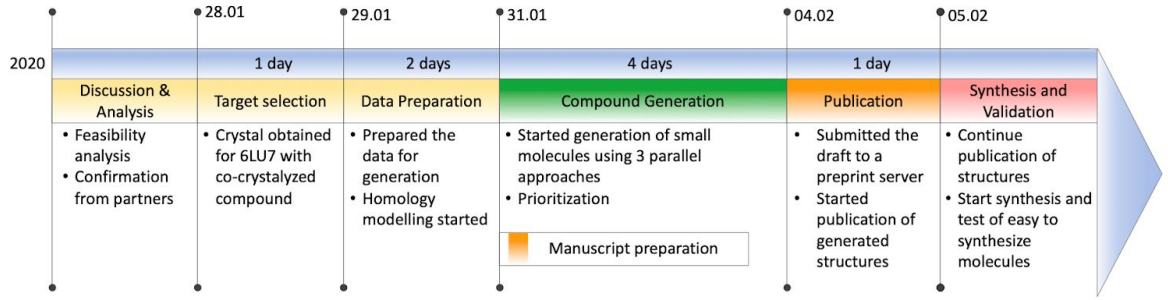

Figure 2: Insilico Medicine 2019-nCoV Small Molecule Generation Sprint Timeline

Input data and datasets

Crystal structure of 2019-nCoV 3C-like protease

The crystal structure of 2019-nCoV 3C-like protease was obtained from Dr. Rao's laboratory. The structure was solved with a 2.1-angstrom resolution in complex with the covalent inhibitor named N3. The SARS-CoV main protease has been previously crystallized with the same inhibitor.[17] The ligand was extracted from the crystal and employed in the ligand-based generation. Then, the binding site was annotated utilizing our proprietary pocket module to create amino acid residues mapping suitable as input data for target structure-based generation.

Homology modelling

The homology model of the 2019-nCoV 3C-like protease in complex with non-covalent ligand was built using the primary sequence corresponding to its crystal structure provided by Dr. Rao's laboratory (*vide supra*). The X-ray structure 4MDS [37] (1.6 Å resolution) of SARS-CoV M^{pro} was used as a template which was co-crystallized with a non-covalent inhibitor and had a very high level of similarity with the 2019-nCoV 3C-like protease (95.25% identity). The homology modelling was performed using SWISS-MODEL [38] [39]. Given the almost complete identity of 2019-nCoV and SARS-CoV proteases in their ligand binding sites, we further refined the obtained homology model with the inhibitor bound in ligand pocket using position restrained minimization with GROMACS [40] with the C_α atoms of protein and all heavy atoms of ligand restrained by harmonic constraints (k_{spring}=100 kJ/mol/nm^2). Two protonation states of His41 situated in the binding pocket were considered. The constructed homology model was preprocessed for generation as described above for crystal structure.

Co-crystalized fragment

The 3D structure of the N3 inhibitor was extracted from the solved complex. The propanoate substructure was replaced by a propenoate, then it was converted to the E-configuration to restore the compound structure that occurred before covalent addition. The obtained conformation was used to build the shape of the ligand as well as two pharmacophore hypotheses using our proprietary modules. For each hypothesis, 7 pharmacophore points were selected according to the interactions in the initial crystal structure and coverage of the peptidomimetic scaffold. The constructed ligand shape and hypotheses were exploited for estimating how generated structures fit the structural features essential for binding.

Protease dataset

The *protease dataset* was assembled with molecules active against various proteases in enzymatic assays extracted from the Integrity database [41], Experimental Pharmacology module and ChEMBL [42,43]. The records from the ChEMBL database were downloaded with the following activity standard types: 'Potency', 'IC_{50}', 'K_i', 'EC_{50}', 'K_d' (assay confidence score ≥ 8, assay type: B, F). The activities from the Integrity database were downloaded using the following parameters: 'IC_{50}', 'K_i', 'EC_{50}', 'K_d', and mass concentrations (e.g. mg/l) were converted to M values by molecular weight. Integrity records were standardized using the pChEMBL value format (logarithmic scale –log10 of a numeric value in M) and merged with the records from ChEMBL. The resulting records with pChEMBL values less than 5.0 (10μM in terms of IC_{50}) were then removed.

The structural duplicates were filtered out after the standardization procedure and the removal of salt parts from salt compositions. Mild medicinal chemistry filters (MCFs) were applied to filter out highly non-drug-like molecules (e.g. metals, polycondensed aromatics, chloramines, radicals, hydrazines, isonitriles, nitroso compounds) as well as structures containing cycles bigger than 8 atoms and polypeptides (n≥4). The resulting dataset contained 60,293 unique structures.

To tailor the scoring and the rewarding functions to the given problem, a *protease peptidomimetics dataset* was collected from the *protease dataset* using SMARTS queries for common peptidomimetic substructures, filtering compounds with pChEMBL value less than 6.0, and suppressing the overrepresented chemotypes. The resulting *protease peptidomimetics dataset* contained 5,891 compounds.

Generative pipeline

We launched Insilico Medicine's generative chemistry platform for every input data type: crystal structure, homology model and co-crystalized ligand.

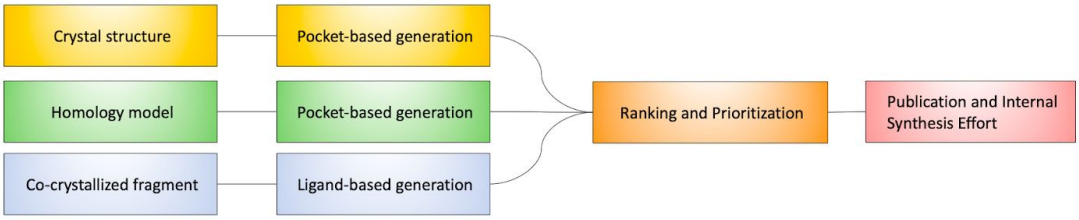

Figure 3: Insilico Medicine 2019-nCoV Small Molecule Generation Procedure

During the generative phase, a total of 28 machine learning (ML) models generated molecular structures and optimized them with reinforcement learning (RL) employing the reward function described below. We used different ML approaches such as generative autoencoders, generative adversarial networks, genetic algorithms, and language models. The models exploited various molecular representations, including fingerprints, string representations, and graphs. Every model was optimizing the reward function to explore the chemical space, exploit promising clusters, and generate new molecules with high scores. The rewarding function was a weighted sum of multiple intermediate rewards: medicinal chemistry and drug-likeness scoring, active chemistry scoring, structural scoring (fitting to ligand features and/or binding pocket), novelty scoring, and diversity scoring.

Medicinal chemistry scoring assigned a low reward to molecules with structural alerts and a high reward to molecules with useful substructures. Drug-likeness scoring drove the generation towards the molecules with molecular properties that are representative for *protease peptidomimetics dataset*—logP: 1.49–6.00; Molecular weights (MW): 400–800; Number of hydrogen bond donors (HBD): 1–10; Number of hydrogen bond acceptors (HBA): 2–10; Topological polar surface area (TopoPSA): 80–210; MCE-18[44]: 40–180; Number of stereocenters (nSC): 0–3.

Active chemistry scoring utilized self-organizing maps trained on *protease peptidomimetics dataset*. We used novelty and diversity scoring in the optimization procedure to explore the chemical space and output a novel and diverse set of molecular structures. Generated compounds were penalized for the similarity to the existing molecules and previously explored clusters. We performed structural scoring with the provided crystal structure or homology model and pharmacophore/shape scoring for structure-based and ligand-based generations,

respectively. We ran the distributed pipeline for 72 hours on the internal computing cluster with 64 NVIDIA Titan V GPUs.

Results

In this study, we used our proprietary generative chemistry pipeline utilizing the knowledge of the crystal structure and homology model of the target protein. We launched the generative pipeline three times for every input data type: crystal structure, homology models and co-crystalized ligand. For each launch, the highest-ranking structures were selected for further analysis. Figure 4 shows some representative examples from the chemical space produced by our generative pipeline launch for the crystal structure. More compounds for generations based on crystal structure, homology models, and co-crystallized ligand are available as described in the section “Availability of structures”. These virtual structures display high 3D-complexity and correspondingly high values of MCE-18, and contain stereo- and/or spiro centers (Table 1), which are common characteristics of peptidomimetics and PPI inhibitors. We assessed the similarity of the structures with compounds from the ChEMBL database using the search engine on the ChEMBL website. The analysis revealed that there are no molecules with the same core structure among the compounds with similarity values more than 0.7 (see Figure 5).

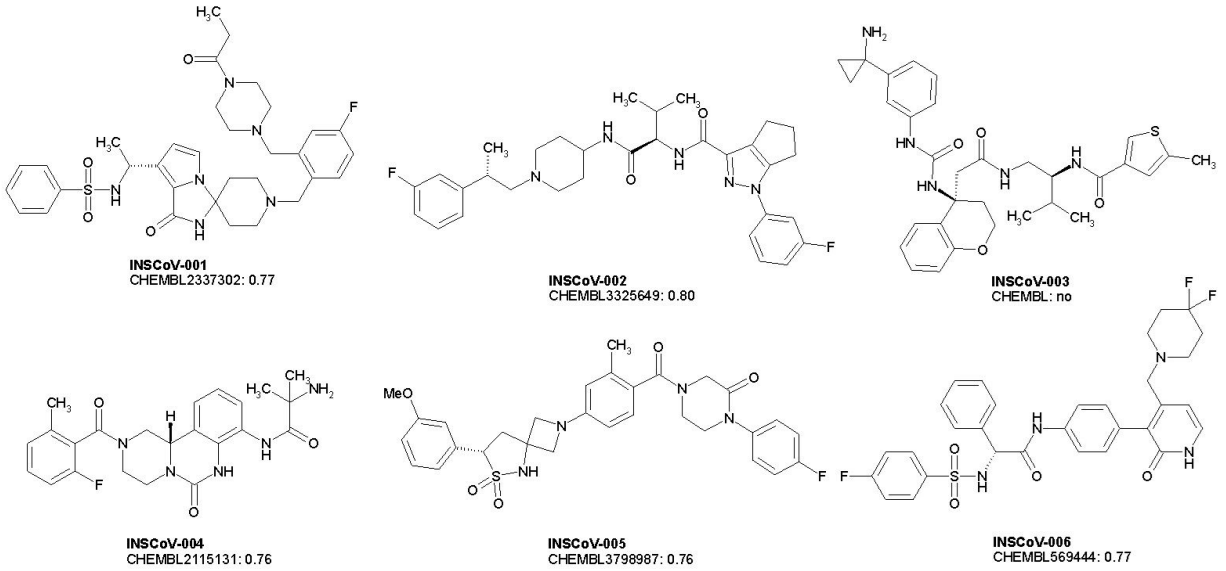

Figure 4. Representative examples of the structures generated to target the main protease of 2019-nCoV. Novelty was assessed using similarity search in ChEMBL Database. ChEMBL ID numbers and maximal similarity coefficients are listed, “no” means that there are no structures with similarity >0.7.

Table 1. The physicochemical descriptors for the representative examples of generated structures. MW—molecular weight, nRot—number of rotatable bonds, nAR—number of aromatic rings, nSC—number of stereocenters, HBA—number of hydrogen bond acceptors, HBD—number of hydrogen bond donors, MCE-18—medicinal chemistry evolution 2018 descriptor.

ID	MW	nRot	nAR	nSC	HBA	HBD	MCE-18	TopoPSA
INSCoV-001	636	9	3	1	4	4	162	115
INSCoV-002	563	9	3	2	3	3	100	79
INSCoV-003	589	10	3	2	4	7	105	163
INSCoV-004	439	3	2	1	3	5	88	108
INSCoV-005	578	5	3	1	5	1	163	108
INSCoV-006	610	9	3	1	4	4	104	120

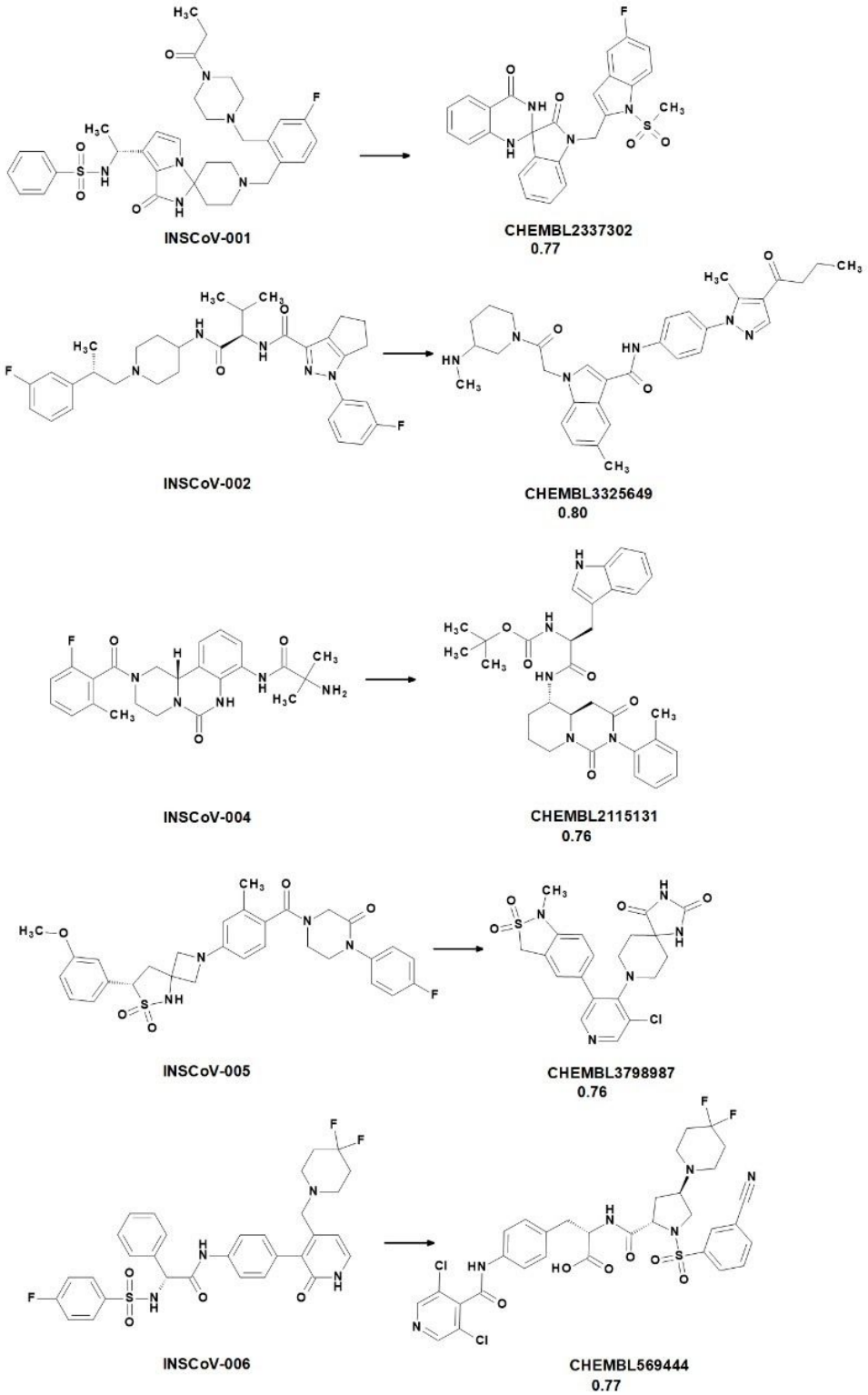

Figure 5. The assessment of similarity between generated structures and compounds from the ChEMBL database utilizing the tool implemented into ChEMBL search. The closest molecules from ChEMBL with ID numbers are presented on the right as well as ChEMBL similarity scores.

Availability of Structures

The most recent data package is available at insilico.com/ncov-sprint. We will continue to update the data package with new compounds during the following weeks. These data could be used to perform subsequent computer modelling simulations or to synthesize and test the compounds *in vitro* against the 2019-nCoV main protease.

Conclusion and discussion

Despite the economic and societal impact of CoV infections and the likelihood of future outbreaks of even more serious pathogenic CoVs in humans, there is still a lack of effective antiviral strategies to treat CoVs and few options to prevent CoV infections [10]. Given the high prevalence and wide distribution of CoVs, the novel virus could emerge periodically in humans as a consequence of frequent cross-species infections and occasional spillover events [45]. The development of effective and time-efficient computational methods for designing compounds that can treat CoV infections is critical. In this study, we have used our integrated AI-based drug discovery pipeline to generate novel drug compounds against 2019-nCoV. The results demonstrate the cost-effectiveness and time efficiency of this type of new method for the development of novel treatments against CoV infections.

Acknowledgments:

The authors would like to thank Dr. Kerry Blanchard for valuable advice, introductions, encouragement, and edits. Dr. Rao's team provided the crystal structure for 6LU7. We would like to thank WuXi AppTec team which graciously agreed to provide full support on the synthesis and biological assay development. We would like to thank Dr. Ding Sheng and Dr. Pan Lirong from GHDDI for providing the compounds-related databases. We would like to thank our board members and especially Nisa Leung for supporting and encouraging this emergency initiative which falls outside of Insilico's scope of commercial efforts. We would like to thank NVIDIA for their generous support and samples of the GPUs.

Conflicts of interest

All authors are affiliated with Insilico Medicine, a company developing an AI-based end-to-end integrated pipeline for drug discovery and development and engaged in aging and cancer research.

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. & Munster, V. J. SARS and MERS: recent insights into emerging coronaviruses. *Nat. Rev. Microbiol.* **14**, 523–534 (2016).
3. Menachery, V. D. *et al.* A SARS-like cluster of circulating bat coronaviruses shows potential for human emergence. *Nat. Med.* **21**, 1508–1513 (2015).
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. Dong, N. *et al.* Genomic and protein structure modelling analysis depicts the origin and infectivity of 2019-nCoV, a new coronavirus which caused a pneumonia outbreak in Wuhan, China. *Microbiology* (2020).
7. Letko, M. & Munster, V. Functional assessment of cell entry and receptor usage for lineage B β-coronaviruses, including 2019-nCoV. *Microbiology* 6117 (2020).
8. Zhao, Y. *et al.* Single-cell RNA expression profiling of ACE2, the putative receptor of Wuhan 2019-nCov. *Bioinformatics* (2020).
9. Ji, W., Wang, W., Zhao, X., Zai, J. & Li, X. Homologous recombination within the spike glycoprotein of the newly identified coronavirus may boost cross-species transmission from snake to human. *J. Med. Virol.* (2020) doi:10.1002/jmv.25682.
10. de Wilde, A. H., Snijder, E. J., Kikkert, M. & van Hemert, M. J. Host Factors in Coronavirus Replication. *Curr. Top. Microbiol. Immunol.* **419**, 1–42 (2018).
11. Xia, B. & Kang, X. Activation and maturation of SARS-CoV main protease. *Protein Cell* **2**, 282–290 (2011).

12. Nukoolkarn, V., Lee, V. S., Malaisree, M., Aruksakulwong, O. & Hannongbua, S. Molecular dynamic simulations analysis of ritonavir and lopinavir as SARS-CoV 3CL(pro) inhibitors. *J. Theor. Biol.* **254**, 861–867 (2008).
13. Fan, K. *et al.* Biosynthesis, purification, and substrate specificity of severe acute respiratory syndrome coronavirus 3C-like proteinase. *J. Biol. Chem.* **279**, 1637–1642 (2004).
14. Thiel, V. *et al.* Mechanisms and enzymes involved in SARS coronavirus genome expression. *J. Gen. Virol.* **84**, 2305–2315 (2003).
15. Goetz, D. H. *et al.* Substrate specificity profiling and identification of a new class of inhibitor for the major protease of the SARS coronavirus. *Biochemistry* **46**, 8744–8752 (2007).
16. Adedeji, A. O. & Sarafianos, S. G. Antiviral drugs specific for coronaviruses in preclinical development. *Curr. Opin. Virol.* **8**, 45–53 (2014).
17. Yang, H. *et al.* Design of wide-spectrum inhibitors targeting coronavirus main proteases. *PLoS Biol.* **3**, e324 (2005).
18. Vanhaelen, Q. *et al.* Design of efficient computational workflows for *in silico* drug repurposing. *Drug Discov. Today* **22**, 210–222 (2017).
19. Karaman, B. & Sippl, W. Computational Drug Repurposing: Current Trends. *Curr. Med. Chem.* **26**, 5389–5409 (2019).
20. Computational Methods for Drug Repurposing. *Methods in Molecular Biology* (2019) doi:10.1007/978-1-4939-8955-3.
21. Mani, D., Wadhwani, A. & Krishnamurthy, P. T. Drug Repurposing in Antiviral Research: A Current Scenario. *J. Young Pharm.* **11**, 117–121 (2019).
22. Dyall, J. *et al.* Repurposing of clinically developed drugs for treatment of Middle East respiratory syndrome coronavirus infection. *Antimicrob. Agents Chemother.* **58**, 4885–4893 (2014).

23. Dyall, J. *et al.* Middle East Respiratory Syndrome and Severe Acute Respiratory Syndrome: Current Therapeutic Options and Potential Targets for Novel Therapies. *Drugs* **77**, 1935–1966 (2017).
24. Li, Y. *et al.* Therapeutic Drugs Targeting 2019-nCoV Main Protease by High-Throughput Screening. *Pharmacology and Toxicology* (2020).
25. Xu, Z. *et al.* Nelfinavir was predicted to be a potential inhibitor of 2019-nCov main protease by an integrative approach combining homology modelling, molecular docking and binding free energy calculation. *Pharmacology and Toxicology* 264 (2020).
26. Liu, X. & Wang, X.-J. Potential inhibitors for 2019-nCoV coronavirus M protease from clinically approved medicines. *Bioinformatics* (2020).
27. Zhavoronkov, A., Vanhaelen, Q. & Oprea, T. I. Will Artificial Intelligence for Drug Discovery Impact Clinical Pharmacology? *Clin. Pharmacol. Ther.* (2020) doi:10.1002/cpt.1795.
28. Kadurin, A. *et al.* The cornucopia of meaningful leads: Applying deep adversarial autoencoders for new molecule development in oncology. *Oncotarget* vol. 8 (2016).
29. Polykovskiy, D. *et al.* Entangled Conditional Adversarial Autoencoder for *de Novo* Drug Discovery. *Mol. Pharm.* **15**, 4398–4405 (2018).
30. Zhavoronkov, A. *et al.* Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nat. Biotechnol.* **37**, 1038–1040 (2019).
31. Kadurin, A., Nikolenko, S., Khrabrov, K., Aliper, A. & Zhavoronkov, A. druGAN: An Advanced Generative Adversarial Autoencoder Model for *de Novo* Generation of New Molecules with Desired Molecular Properties *in Silico*. *Mol. Pharm.* **14**, 3098–3104 (2017).
32. Putin, E. *et al.* Adversarial Threshold Neural Computer for Molecular *de Novo* Design. *Mol. Pharm.* **15**, 4386–4397 (2018).
33. Putin, E. *et al.* Reinforced Adversarial Neural Computer for *de Novo* Molecular Design. *J.*

Chem. Inf. Model. **58**, 1194–1204 (2018).

34. Kuzminykh, D. *et al.* 3D Molecular Representations Based on the Wave Transform for Convolutional Neural Networks. *Mol. Pharm.* **15**, 4378–4385 (2018).
35. Aliper, A. *et al.* Deep Learning Applications for Predicting Pharmacological Properties of Drugs and Drug Repurposing Using Transcriptomic Data. *Mol. Pharm.* **13**, 2524–2530 (2016).
36. Zhavoronkov, A. Artificial Intelligence for Drug Discovery, Biomarker Development, and Generation of Novel Chemistry. *Mol. Pharm.* **15**, 4311–4313 (2018).
37. Turlington, M. *et al.* Discovery of N-(benzo[1,2,3]triazol-1-yl)-N-(benzyl)acetamido)phenyl) carboxamides as severe acute respiratory syndrome coronavirus (SARS-CoV) 3CLpro inhibitors: identification of ML300 and noncovalent nanomolar inhibitors with an induced-fit binding. *Bioorg. Med. Chem. Lett.* **23**, 6172–6177 (2013).
38. Waterhouse, A. *et al.* SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res.* **46**, W296–W303 (2018).
39. Guex, N., Peitsch, M. C. & Schwede, T. Automated comparative protein structure modeling with SWISS-MODEL and Swiss-PdbViewer: a historical perspective. *Electrophoresis* **30 Suppl 1**, S162–73 (2009).
40. Abraham, M. J. *et al.* GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1-2**, 19–25 (2015).
41. Clarivate Analytics Integrity. https://integrity.clarivate.com/integrity/.
42. Gaulton, A. *et al.* ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **40**, D1100–7 (2012).
43. *CHEMBL database release 25*. http://ftp.ebi.ac.uk/pub/databases/chembl/ChEMBLdb/releases/chembl_25 (2019)

doi:10.6019/CHEMBL.database.25.

44. Ivanenkov, Y. A., Zagribelnyy, B. A. & Aladinskiy, V. A. Are We Opening the Door to a New Era of Medicinal Chemistry or Being Collapsed to a Chemical Singularity? *J. Med. Chem.* **62**, 10026–10043 (2019).
45. Zhu, N. *et al.* A Novel Coronavirus from Patients with Pneumonia in China, 2019. *N. Engl. J. Med.* (2020) doi:10.1056/NEJMoa2001017.