Team name	COVID19DDC (COVID-19 Drug Discovery Consortium)
Team members	Stan Watowich; Suman Sirimulla; William Allen; Xiaodong Cheng; Robert Davey; Andrea Dimet; Francisco Enguita; Amit Gupta; Yurii Moroz; Pei-Yong Shi; Clifford Stephan; Adrian Varela-Alvarez; Jin Wang; Mark White
Affiliations	University of Texas Medical Branch; Univeristy of Texas El Paso; Texas Advanced Computer Center; University of Texas Health Science Center; Boston University; Universidade de Lisboa; Texas Medical Center Innovation Institute; Enamine, Ltd; Galveston National Laboratory; Texas A&M University Health Science Center; Baylor College of Medicine
Contact email	watowich@xray.utmb.edu
Contact phone number	cell: (01) 832-613-5356
Protein targets	- nsp3; ADP ribose phosphatase (MAC1 or X domain) - nsp5; Main protease (Mpro; also called 3CL-pro) - nsp15; RNAase - nsp16 / nsp10; meethyltransferase - nsp3 domain; papain-like_protease (PLpro) - nsp12; RNA dependent RNA polymerase (RdRp)

Section 1: methods & metrics
Describe what methods you have used, how they are independent from one another, what your workflow was, how you performed the cross-correlation between your methods. If applicable, please report estimated performance metrics of your methods, such as accuracy, sensitivity, false-discovery rate, etc., and how those metrics were obtained (e.g. cross-validation). Please provide key references if available.

Overview. The COVID19 Drug Discovery Consortium (COVID19DDC), an international team of leading computational scientists, medicinal chemists, biochemists, and virologists, was assembled April 6, 2020 to rapidly identify promising new COVID-19 antivirals that function through varied mechanisms-of-action. This project combines massive virtual screening calculations with cytotoxicity and cell-based SARS-CoV-2 antiviral assays to accomplish this goal. The most potent antiviral compounds emerging from this project will transition to preclinical development, including GMP manufacturing and GLP safety studies in preparation for first-in human clinical studies.

Phase 1. Virtual screening. The Consortium secured priority access to computational resources within the Texas Advanced Computing Center (TACC; University of Texas, Austin, TX), including several of the world's most powerful supercomputers, to complete the massive virtual screening calculations required for this pr ject. These calculations leverage the DrugDiscovery@TACC platform, a globally-accessible open-computing resource previously developed by Dr. Watowich in collaboration with TACC computer scientists. Additionally, BOINC@TACC volunteer computing systems are being enlisted to accelerate this effort.

Six different SARS-CoV-2 proteins (see above list) were prioritized as small molecule drug targets and their 3-dimensional crystallographic structures retrieved from the Protein Data Bank. For each protein chain in the available unique PDB structures, active sites conducive for inhibitor binding were identified. For example, two PDB (Protein Data Bank) entries (6W02, 6W6Y) were available for the SARS-CoV-2 nsp3 ADP ribose phosphatase protein, and each PDB file contained two structures in the crystallographic asymmetric unit, which resulted in four unique structures for use in the virtual screening calculations. All unique protein chains (a total of 17 different structures) for each SARS-CoV-2 protein target, together with chemical

libr ries containing 2.6 million drug-like, commercially-available small molecules, were optimized and parameterized for very large-scale virtual screening. Scripts were streamlined to efficiently perform large-scale Vina-based virtual screening calculations on TACC supercomputing systems (Frontera, Stampede2, Lonestar5, and BOINC@TACC).

Vina-based virtual screening calculations against the SARS-CoV-2 proteins were rapidly completed. The virtual screening phase of this project systematically calculated optimal binding structures and energies between each small molecule in a 2.6 million-member library (encompassing Enamine Ltd's HTS, Advanced, and Premium chemical collections) and each SARS-CoV-2 non-structural protein (e.g., Mpro, RdRp, helicase, RNAase, papain-like protease). These calculations identified several thousand small molecules ("hits") predicted to tightly bind and inhibit critical active sites within each of the SARS-CoV-19 non-structural proteins. This list of several thousand hits for each protein target was subjected to a second round of virtual screening using the Schrodinger Glide algorithm. Results from the Vina and Glide scoring calculations were re-ordered using ordinal averaging to produce a consensus list of compounds ranked by average consensus ordinal score.

Phase 2. Assemble Prospective COVID-19 Antiviral Library. For each PDB chain associated with the six unique SARS-CoV-2 non-structural proteins, approximately 100 high-scoring hits were selected from the virtual screening phase. These hits were filtered to remove compounds that failed a subset of PAINS (pan-assay interference compounds) alerts recently suggested as reliable and predictive for identifying assay interfering compounds; this filter provided a balance between prematurely eliminating potent compounds and testing compounds that interfere with planned cytotoxicity and antiviral assays. The resulting hits were further evaluated based on physiochemical properties, ADMET (absorption, distribution, metabolism, excretion, toxicity) profiles, chemical diversity, molecular target, and intermolecular binding interactions. Following ADMET/PAINS filtering, one hundred top-ranked hits were selected for each SARS-CoV-2 non-structural proteins, with similar numbers of hits chosen for each of the PDB chains. These 600 hits, targeting equally the six different SARS-CoV-2 enzymes, comprised our initial Prospective COVID-19 Antiviral Library. Each compound in this library was ordered from Enamine's stockpiled chemical collections, verified for purity and chemical composition, and assembled onto 384-well plates (Labcyte Echo Qualified 384-Well Low Dead Volume Microplates; 300 compounds/plate) as 10 mM solutions in DMSO. This plate format was adopted for the Prospective COVID-19 Antiviral Library to maintain compatiblity with the liquid-handling robotic systems used in our *in vitro* high-throughput cytotoxicity and antiviral assays. These assays, described below, were performed concurrently to accelerate the project and at separate locations to avoid taxing high-containment resources needed for antiviral screening. Note: compounds were ordered May 13, 2020 and arrived at our US-based testing facilities on May 29, 2020.

Phase 3. Cytotoxicity screening of the Prospective COVID-19 Antiviral Library. Cytotoxicity assays were performed at the Texas A&M University Institute for Biosciences & Technology (Houston, TX, USA) to establish CC_{50} (i.e., concentration associated with 50% reduction of cell viability) values for each small molecule within our Prospective COVID-19 Antiviral Library. Drs. Peter Davies and Clifford Stephan, co-directors of the Combinatorial Drug Discovery High-Throughput Screening (HTS) facility (Texas A&M Institute of Biosciences & Technology), have the necessary cutting-edge robotic and high-content imaging resources to rapidly perform the critical cytotoxicity measurements in Vero cells. Cytotoxicity screening was performed using duplicate technical and biological replicates, with each experiment testing 6-concentrations of each library compound spanning 0.1 uM to 50 uM to establish robust dose-response curves and accurate CC_{50} values. These assays were completed June 10, 2020.

Phase 4. Antiviral screening of the Prospective COVID-19 Antiviral Library. Cell-based antiviral assays were performed to establish $\log EC_{50}$ values (the effective concentration required to exponentially reduce SARS-CoV-2 replication titers) for each library compound. Renowned infectious disease collaborators at two ma-

jor US National Biocontainment Laboratories have the resources and expertise to test for SAR-CoV-2 antiviral activity in cell culture. Dr. Robert Davey (National Emerging Infectious Diseases Laboratories [NEIDL], Boston University, Boston, MA) utilized their BSL4-based robotic technologies and high-content imaging resources to screen the Prospective COVID-19 Antiviral Library for compounds that prevent SARS-CoV-2 replication. Antiviral assays were performed in duplicate in Vero cells with 4-concentrations (spanning 1 uM to 50 uM) of library compounds to establish preliminary dose-response curves and $logEC_{50}$ values suitable to identify promising antivirals. These assays were completed the week of June 22, 2020.

Approximately 20 of the most promising (i.e., largest CC_{50}/$logEC_{50}$ values) antiviral compounds identified by Drs. Stephan's and Davey's teams will be delivered to Dr. Pei-Yong Shi at the Galveston National Laboratory (GNL; University of Texas Medical Branch, Galveston, TX) for detailed characterization and independent validation. Comprehensive dose-ranging antiviral studies (spanning compound concentrations from 10 nM to 50 uM) will utilize Dr. Shi's recently developed novel GFP-expressing SARS-CoV-2 assay and measure virus replication in several cell lines (e.g., Vero, Calu3). These studies are best suited for the GNL since they do not require NEIDL's high-value HTS robotic resources and take advantage of the numerous virus-permissive cell lines maintained within the GNL's BSL3 laboratories. The resulting robust dose-response curves generated by Dr. Shi's team will provide accurate $logEC_{50}$ values in multiple cell lines for each novel COVID-19 antiviral. These validating efficacy assays are targeted for completion by mid-July, 2020.

Section 2: targets
Describe for each protein target: why you chose it, from which source you obtained it (e.g., insidecorona.net / covid.molssi.org / rcsb.org) and why this is the best quality structure, if any preprocessing (e.g., energy minimization, residue correction, alternative folding, …) was performed.

Target 1: SARS-CoV-2 nsp3 (ADP ribose phosphatase; MAC1 or X domain) is believed responsible for decreased sensitivity to interferon-alpha. This protein also contains de-mono-ADP-ribosylation activity that is critical for virus function. nsp3 is thought to play a role in supressing host innate immune gene expression, such as IL-6 and IFN-beta, in part through mono-ADP-ribosylation.

Atomic structures were downloaded from the Protein Data Bank (PDB) as file 6W02.pdb, which contained two independent monomors (chain A, B) of nsp3 (ADP ribose phosphatase; MAC1 or X domain) each in complex with ADP ribose (APR). Resolution of the X-ray crystallographic structure is 1.5 Å. Chain B was superimposed onto chain A to position all structures in an identical reference frame. Test docking of APR to the 6W02 structures (using a 26x26x26 Å box centered at {3, -3.8, -21}) using the Vina program generated a docked pose that was superposable on the cocrystal structure.

Atomic structures were downloaded from Protein Data Bank as file 6W6Y.pdb, which contained two independent monomors (chain A, B) of nsp3 (ADP ribose phosphatase; MAC1 or X domain) each in complex with APR. Resolution of the X-ray crystallographic structure is 1.45 Å. Chains A and B were superimposed onto chain A of structure 6W02 to position all structures in an identical reference frame. Test docking of APR to the 6W02 structures (using a 26x26x26 Å box centered at {3, -3.8, -21}) using the Vina program generated a docked pose that was superposable on the cocrystal structure.

Target 2: SARS-CoV-2 nsp5 (Main protease Mpro; 3CL-pro) is the "main protease" of the virus responsible for the expression of the two most conserved replicase domains in nidoviruses (the order of coronoviruses), namely RdRp and RNA helicase.

An atomic structure was downloaded from the Protein Data Bank as 6Y84.pdb. This structure is a 1.39 Å resolution apo structure and contains 1 chain in the asymmetric unit. There are many multiple alternative sidechain conformations present in the PDB file. ; although only residues Ala116, Cys117, Lys137, Met165

were within 10 Å of catalytic site. Conformers Ala116A, Cys117A, and Lys137B were used for virtual screening studies, since these sidechain orientations overlapped the PDB 5R80 cocrystal structure. However, the Met165 residue orientations in 6Y84.pdb were distinct from the Met165 sidechain orientation noted in the 5R80 cocrystal structure. Thus, both conformations of Met165 contained in 6Y84.pdb were used in virtual screening as independent structures.

An atomic structure was downloaded from the Protein Data Bank as 5R80.pdb. This structure is a 1.93 Å resolution structure of the protein bound to a hydrolase inhibitor (RZG) identified from fragment-based screening. The PDB structure contains 1 chain in the asymmetric unit. Test docking of ZRG to the 5R80 structure (using a 28x28x28 Å box centered at {8.5, -3.6, 18.7}) using the Vina program generated a docked pose that largely reproduced the cocrystal structure.

Target 3: *SARS-CoV-2* nsp15 is an RNA endoribonuclease which may influence cellular RNA processing or coronavirus replication or transcription; there is some evidence that coronaviruses may require nuclease activity in order to synthesize RNA.

An atomic structure was downloaded from the Protein Data Bank as 6W01.pdb. This is a 1.93 Å resolution structure of the endoribonuclease in complex with a citrate molecule. The PDB file contains two monomers (chains A, B) in the asymmetric unit. Chain B was superimposed onto chain A to position both structures in an identical reference frame.The catalytic center is located at His235, His250, and Lys290. Vina-based virtual screening used a 28x28x28 Å box centered on the NE2 atom of His250.

Target 4: *SARS-CoV-2* nsp16/nsp10 forms the methyltransferase-stimulatory factor complex of the virus. nsp16 is an S-adenosylmethionine (SAM)-dependent (nucleoside-2'-O)-methyltransferase, whose full activity requires the activating partner nsp10. Both NSP 10 and NSP16 have been deemed essential for a functional replicase-transcriptase complex

An atomic structure was downloaded from the Protein Data Bank as 6W01.pdb. This 1.93 Å resolution structure contains the endoribonuclease in complex with a citrate molecule. The PDB file contains two monomers (chains A, B) in the asymmetric unit. Chain B was superimposed onto chain A to position both structures in an identical reference frame. The catalytic center is located at His235, His250, and Lys290. Vina-based virtual screening used a 28x28x28 Å box centered on the NE2 atom of His250.

Target 5: *SARS-CoV-2* nsp3 domain (PLpro) is a papain-like protease required by SARS coronavirus to process its replicase polyproteins.

An atomic structure was downloaded from the Protein Data Bank as 6W9C.pdb. This apo structure is at 2.7 Å resolution, with the asymmetric unit consisting of a trimer (chains A, B, C). Chains B and C were superimposed onto chain A to position all structures in an identical reference frame.

Test docking of a PLpro inhibitor (GRN; extracted from the cocrystal structure 3MJ5) to chain A of 6W9C (using a 24x24x24 Å box centered at {-14, 46, -39}) using the Vina program generated a docked pose that reproduced the cocrystal structure observed in 3MJ5.

Target 6: *SARS-CoV-2* nsp12 is a RNA-dependent RNA polymerase (RdRp) required by SARS coronavirus to replicate.

Atomic structures 6M71.pdb and 7BTF.pdb were downloaded from the Protein Data Bank. These structures are at 2.9 Å and 2.95 Å resolution, respectively. The structures were solved using cryoEM techniques. The observed RNA-dependent RNA polymerases are in complex with cofactors nsp7/nsp8.

Atomic structure 7BV2.pdb was downloaded from the Protein Data Bank. This cryoEM structure is at 2.5 Å resolution, and is complexed with the inhibitor remdesivir. The monomers in 7BTF.pdb and 7BV2.pdb were superimposed onto chain A of 6M71.pdb to position all structures in an identical reference frame.

Related RdRp structures that contained non-nucleotide inhibitors were examined. PDB structures 6QCV, 2JC1, 5QJ1, 5F3T, 5TRJ all have inhibitors that bind at residues 812-815. The structure 4NRU has was solved with an inhibitor bound at Arg836, which is spatially near residue 812. The structure 3UPF had an inhibitor that bound at Arg553, while structure 5IQ6 had an inhibitor bound at Lys500. A virtual screening box with dimensions 30x30x30 and centered at coordinates {114, 121, 125} was sufficiently large to encompass all locations occupied by the above RdRp inhibitors and the remdesivir position located in PDB 7BV2.pdb. Test docking of remdesivir to the 6M71, 7BTF, and 7BV2 structures using the Vina program generated docked poses that poorly reproduced the 7BV2 cocrystal structure, suggesting virtual screening to these targets will be challenging.

In total, 17 different protein atomic structures, spanning 6 unique SARS-CoV-2 enzymes, were used as targets for our virtual screening campaigns.

Section 3: libraries
Describe which libraries you have used, how they were combined, if any compounds were removed / added, why additions are relevant, any unique features of your library, etc. Please provide the sources you obtained the libraries from (if publicly available). Describe the procedure of data preparation (removal of duplicates, standardization, etc). Indicate if different libraries were used for different targets, and why. If possible, provide a download link to your version of the library.

As noted above, our goal was to rapidly complete a bold virtual screening program to identify small molecule "hits" that would be immediately available for testing in antiviral and cytotoxicity assays. Thus, we focused on screening libraries that contained small molecules that were readily available for purchase at high purity. Enamine, Ltd was identified as a global chemical supply source that maintained the world's largest collection of compounds for immediate biological screening, with physical stocks of ~2.6 million validated small molecules in quantities of ~150 mg. Moreover, Enamine could rapidly assemble hundreds to thousands of thee compounds as DMSO stocks in 96-well and 384-well plate formats suitable for robotic liquid handling and high-throughput screening. Enamine provided files for each chemical library (listed below) that contained the description of each molecule in SDF format. In-house software was used to convert SFD-formatted molecular descriptions to PDBQT-formatted files for use in the Vina docking program. The conversion process also identified molecules with chiral centers, generating appropriate stereoisomers for each center. Thus, although the three libraries described below contain 2,647,089 distinct small molecules, a total of ~5.5M unique stereoisomers were screened against each of the 17 protein targets

Library 1: The Enamine Premium Collection contains 43,525 compounds having favorable physicochemical properties (e.g., high Fsp^3, low LogP, low MW). These novel compounds were synthesized in-house by Enamine using their robust and sustainable chemical methodology that leverages their extensive building block inventory. New molecular scaffolds in this library were selected for physicochemical properties, pharmacophore composition, and the cost factor related to synthesis scale-up of promising leads.

Library 2: The Enamine Advanced Collection contains 482,182 unique compounds that have lead-like properties (MW ≤ 350, cLogP ≤ 3, rotB ≤ 7) and/or valuable pharmacophores such carboxylic, primary amino and amide groups. This screening collection has been used extensively as a source of compounds for targeted library design. Multiple functional groups and lead-like properties of compounds in this library offer

valuable avenues for lead optimization. All compounds are checked with Enamine's in-house med-chem filters.

Library 3: The Enamine HTS Collection contains 2,121,382 diverse screening compounds. This collection encompasses versatile chemotypes developed during the past two decades of chemical research at Enamine and its partner academic organizations. These compounds frequently have unusual structures and unique properties. The collection provides a useful resource to examine highly diverse molecules that may require optimization before advancing as clinical candidates.

Section 4: results
Briefly describe you key findings, any interesting trends in your data, a description of your top 5 compounds for each target. If possible, provide a link to a code and/or data repository. Please do not submit randomly selected compounds!

Results: Consensus lists containing the top 100 hits for each SARS-CoV-2 enzyme target are attached using the templates provided. These compounds have completed cytotoxicity evaluation in Vero cells, with only 26 compounds out of the six hundred tested showing modest cytotoxicity. These results will be updated shortly once we have processed and reviewed the antiviral activity data.

Other comments: