

Team name	Way2Drug-IBMC
Team member(s) (firstname lastname; ...)	Vladimir Poroikov; Dmitry Filimonov; Dmitry Druzhilovskiy; Alexander Veselovsky; Alexey Lagunin; Vladlen Skvortsov, Anastasia Rudik; Alexander Dmitriev; Pavel Pogodin; Leonid Stolbov; Olga Tarasova; Sergey Ivanov; Boris Sobolev; Dmitry Karasev; Tatyana Glorizova; Kirill Shcherbakov, Polina Savosina; Nikita Ionov; Nadezhda Biziukova; Vladislav Sukhachev
Affiliation	Institute of Biomedical Chemistry
Contact email	vladimir.poroikov@ibmc.msk.ru
Contact phone number (optional)	+7 916 595-81-05 cell; +7 499 246-09-20 office
Protein targets (for example: 3CLpro/Nsp5, BoAT1, Fc Receptor, Furin, IL6R, M protein, NspX, OrfXx, N, E, etc...) 3 required	3CLpro, RdRp, PLpro, TMPRSS2

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.

Methods: Taking into account the incomplete and sometimes contradictory information about SARS-CoV-2 virus and its interaction with the host cell, which was available on May 6th, 2020 when the project began, we decided to apply the following approach for virtual screening of anticoronaviral hits in the big chemical libraries.

Our approach includes three sequential stages:

- (1) Selection of the potential hits among the 1+ billion compounds based on the similarity assessment using as the “reference drugs” molecules with experimentally determined anticoronaviral action.
- (2) Further filtration and ranking of the selected hits using machine-learning methods implemented in our software PASS and GUSAR (the training sets were extended and improved permanently in the framework of the whole project).
- (3) Verification of the representative examples of the selected compounds using the molecular modeling approach.

General workflow for selection of hits with potential anti-SARS-CoV-2 activity is given in Figure 1.

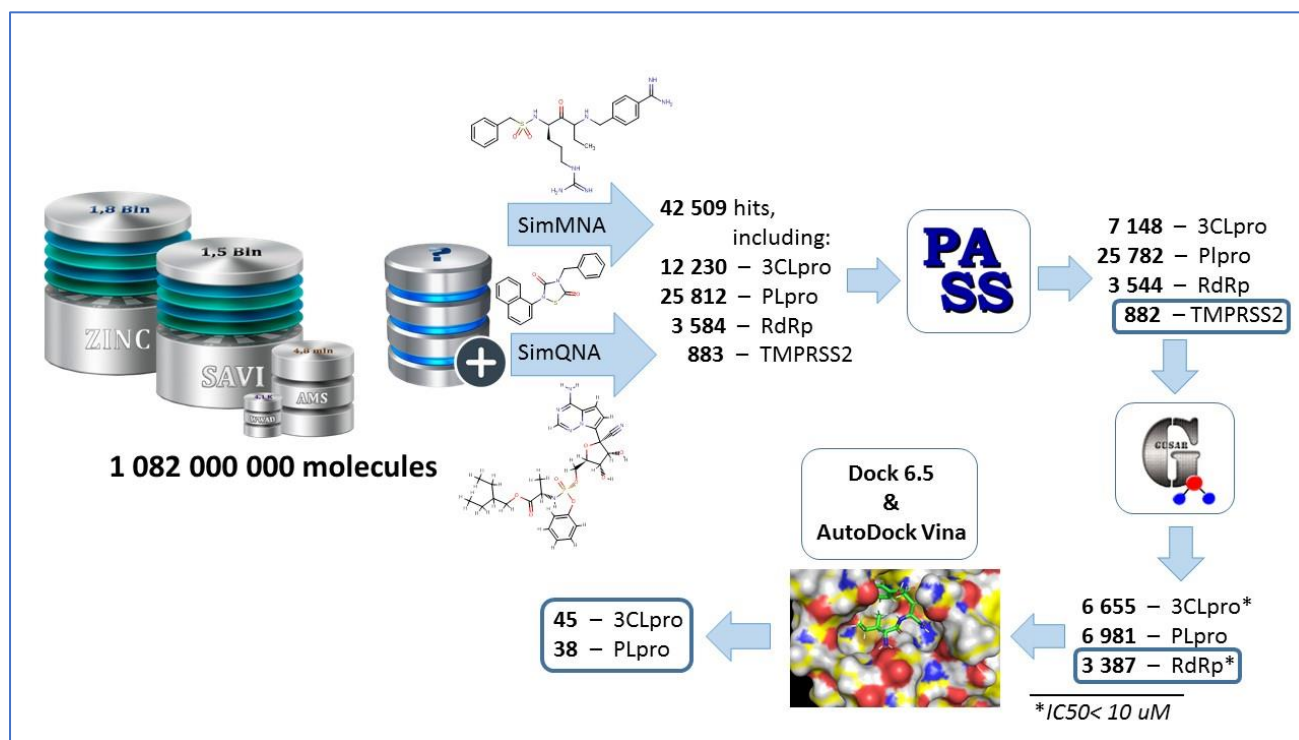


Figure 1. General workflow and results of selection of anti-SARS-CoV-2 hits.

Four different methods applied at these three stages are described below.

1. Similarity assessment.

«Similar molecules exert similar biological activities» [1]. Despite the occasionally observed violation of this suggestion in case of the so-called activity cliffs [2], it is widely used in medicinal chemistry for design of analogues putatively active against the same target or exhibiting the same pharmacological effect [3]. Moreover, it is the “method-of-the-choice” in case of novel pharmacological targets when the number of known molecules interacting with the targets is too small for generating the pharmacophore or (Q)SAR model.

There is no universal method for assessing the similarity between the molecules belonging to different chemical classes and having various biological activities [4, 5]. In the framework of the JEDI COVID-19 Challenge, we decided to use the similarity estimates based on our own descriptors named Multilevel Neighborhoods of Atoms (MNA) [6] and Quantitative Neighborhoods of Atoms (QNA) [7]. These descriptors are successfully applied for analysis of structure-activity relationships for heterogeneous datasets, which completely corresponds to the task of the current project: virtual screening of hits with potential anticoronaviral action among 1+ billion molecules. Recently, we have investigated their applicability to the assessment of activity by similarity for the 16,770 inhibitors of HIV-1 protease, reverse transcriptase and integrase [8], and revealed the possibilities and limitations of this approach [9].

To perform the similarity search and selection of hits with the required biological activity from 1+ billion molecules, we identified the “reference substances” (the most active inhibitors of the four studied targets known in June 2020), which were used as queries. The following reference substances were used.

3CLpro: Five most active compounds were collected from different sources and tested under different experimental protocol. GC376, Tideglusib, 11b, TZDZ-8 activities were taken from the corresponding original publications [34-37]. MAT-POS-916a2c5a-1 was selected from PostEra resource [10]. All of five compounds were tested using SARS-CoV-2 recombinant main protease and showed low micromolar activities.

PLpro: 6-thioguanine, GRL0617, 679818, Psoralidin were taken from the corresponding original publications [11-13] as most active inhibitors of SARS-CoV Papain-like protease.

RDRP: Selection of the most active compounds was carried out in Stanford Coronavirus Antiviral Research Database [14], three chemical compounds were selected, their IDs in widely used databases and common names are: PubChem_CID: 44468216 (GS-441524), PubChem_CID: 121304016 (Remdesivir), ChEMBL_ID: ChEMBL2178720 (Beta-D-N4-Hydroxycytidine). Activity of the GS-441524 and Remdesivir was reported in several preprints [15-21]. The data on activity of the Beta-D-N4-Hydroxycytidine originates from single preprint [22]. All three compounds demonstrated submicromolar activity (EC50) in the tests conducted using SARS-COV-2 and human cell lines to measure antiviral activity. Ability of Remdesivir and GS-441524 to suppress the expression of viral RNA was also studied in addition to the general antiviral effect and compounds achieved submicromolar EC50 values.

TMPRSS2: Selection of the most active compounds was carried out in ChEMBL database [23]. Three chemical compounds having submicromolar Ki values were found: ChEMBL1809250, ChEMBL1229259 and ChEMBL1809251. According to the assay description from ChEMBL, compounds were tested against the recombinant catalytic domain of TMPRSS2 expressed in Escherichia coli using D-cyclohexylalanine-Pro-Arg-AMC as substrate by fluorescence plate reader analysis. Results were published in the paper [24].

In addition to the reference substances found in the available publications and databases, in similarity search we decided to use the structural formulae of ligands complexed with SARS-CoV-2 proteins from Protein Data Bank (PDB) [25]. Information about these molecules is given below.

The PDB database contains over 200 coronavirus protein structures. A small number of structures are crystallized with ligands. To search by similarity, ligands from six complexes were selected (Table 1).

Table 1. Coronavirus target proteins and their ligands obtained from the PDB database.

PDB ID	Ligand ID (name)	Molecular Mass of ligand (Da)	Resolution (Å)	Protein-target	Reference (DOI)
6LU7	PRD_002214 (inhibitor N3)	680	2.2	3CLpro	10.1038/s41586-020-2223-y
7BRP	HU5 (Boceprevir)	519	1.8	3CLpro	to be published
7BRR	K36	485	1.4	3CLpro	to be published
6Y2G	O6K	595	2.2	3CLpro	10.1126/science.abb3405
6W63	X77	459	2.1	3CLpro	to be published
7BV2	F86	371	2.5	RDRP	10.1126/science.abc1560

2. Machine learning with computer program PASS

PASS (Prediction of the Activity Spectra for Substances) [26] is the software that predicts over five thousand biological activities with average accuracy about 96% on the basis of structural formula of drug-like compound [27]. PASS development has been started in the late eighties of the XX century [28], and during the past thirty years we permanently updated the training sets, extended the list of the predictable biological activities, compared the performance of thousands chemical descriptors and dozens mathematical methods. Current version of PASS is based on the analysis of structure-activity relationships (SAR) for 1,025,468 biologically active compounds using MNA descriptors [6] and modified naive Bayes classifier [29]. This method not only allows one to carry out high-accuracy SAR analysis for compounds from the training set, but is also robust enough to provide reasonable estimates of the biological activity spectra of new compounds despite the incompleteness of information in the training set [30]. For a new compound PASS estimates two probabilities: P_a that is the probability of belonging to the subset of “actives”, and P_i that is the probability of belonging to the subset of “inactives”. By default, all compounds, for which estimated $P_a > P_i$, are considered as belonging to the class of “actives”.

PASS performance supersedes those of other known methods for prediction of biological activity profiles, which was shown in the comparative computational experiments [31-33].

PASS Professional version allows creating new training sets, re-training the program to obtain new SAR knowledgebase, and validating the accuracy and predictivity using leave-one-out and 20-fold cross-validation, respectively.

In the framework of the current project, we prepared a specialized training set by collection of information from freely and commercially available databases [23, 24] as well as from many relevant publications. As a cutoff for active compounds we used $IC_{50} < 10 \mu M$. To increase the coverage of the chemical space, all new data regarding the structure and activity of anticoronaviral agents were added to the PASS 2019 training set. After the training and validation, SAR knowledgebase with the following characteristics was obtained:

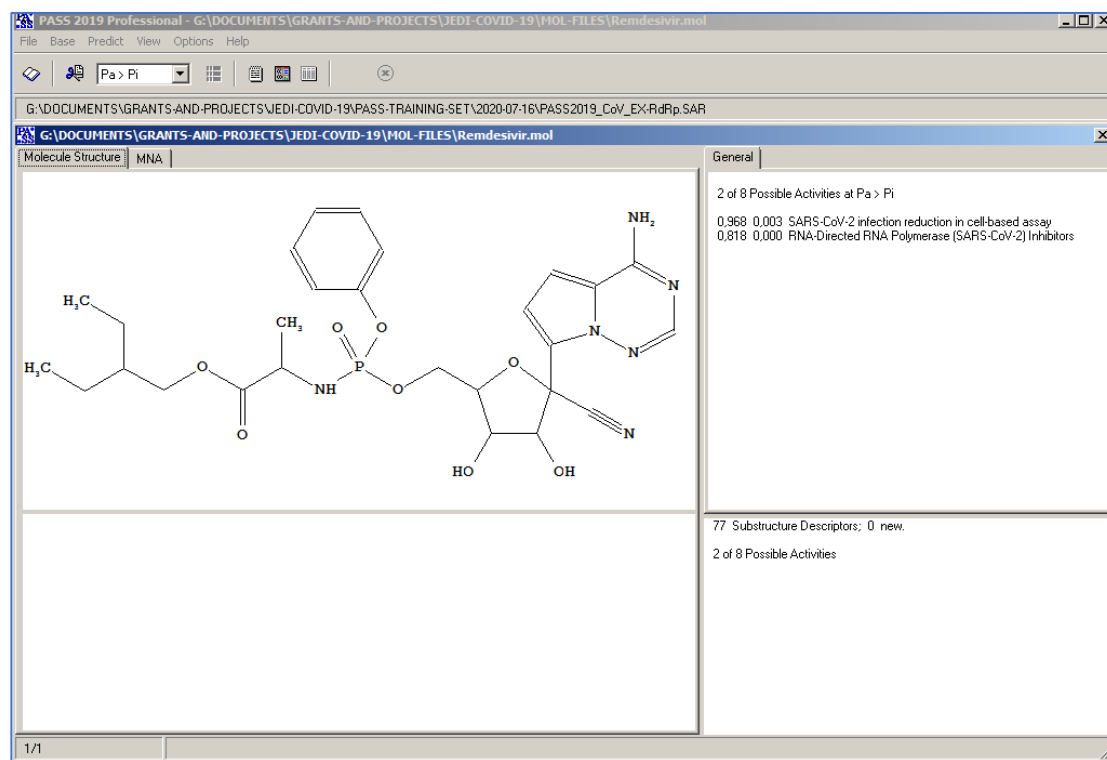
1025630	Substances
106828	Descriptors
8065	Activity Types
8	Selected Activity Types
0.9138	Average IAP

No	Number	IAP	20-Fold	Activity Type
1	62	0.9585	0.9587	3C-Like Protease (SARS-CoV) Inhibitors
2	18	0.9908	0.9909	3C-Like Protease (SARS-CoV-2) Inhibitors
3	6	0.8296	0.8320	Papain-Like Protease (SARS-CoV-2) Inhibitors
4	3	0.9970	0.9980	RNA-Directed RNA Polymerase (SARS-CoV-2) Inhibitors
5	808	0.7535	0.7535	SARS-CoV-2 infection reduction in cell-based assay
6	5	0.9678	0.9684	SARS-CoV-2 viral Entry Inhibitors
7	371	0.8129	0.8147	Spike Glycoprotein (S) (SARS-CoV-2)/ACE2 Interaction Inhibitors
8	3	1.0000	1.0000	Transmembrane Protease Serine 2 (TMPRSS2) Inhibitors

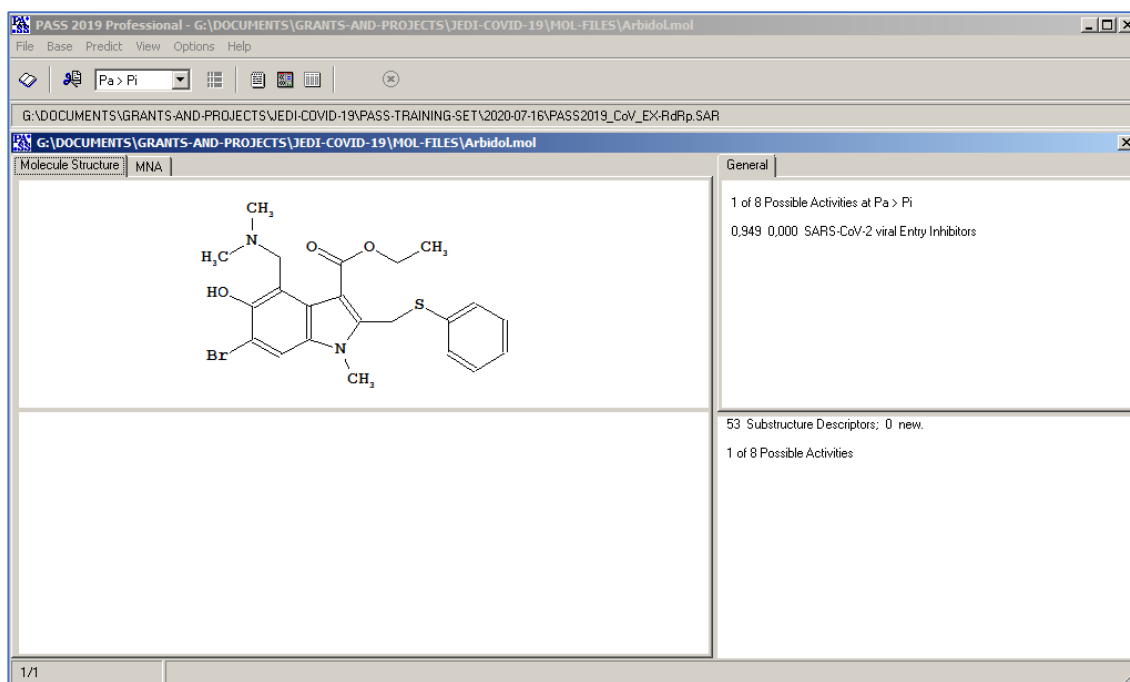
Here: *Number* is the amount of compounds with the particular activity in the training set; IAP is the Invariant Accuracy of Prediction estimated in leave-one-out cross-validation, which is equivalent to the AUC ROC values; 20-fold is the IAP estimate obtained in 20-fold cross-validation.

As one may see from the data presented above, the accuracy (leave one out cross-validation) and predictivity (20-fold cross-validation) of the obtained specialized version of PASS is good enough for its practical application.

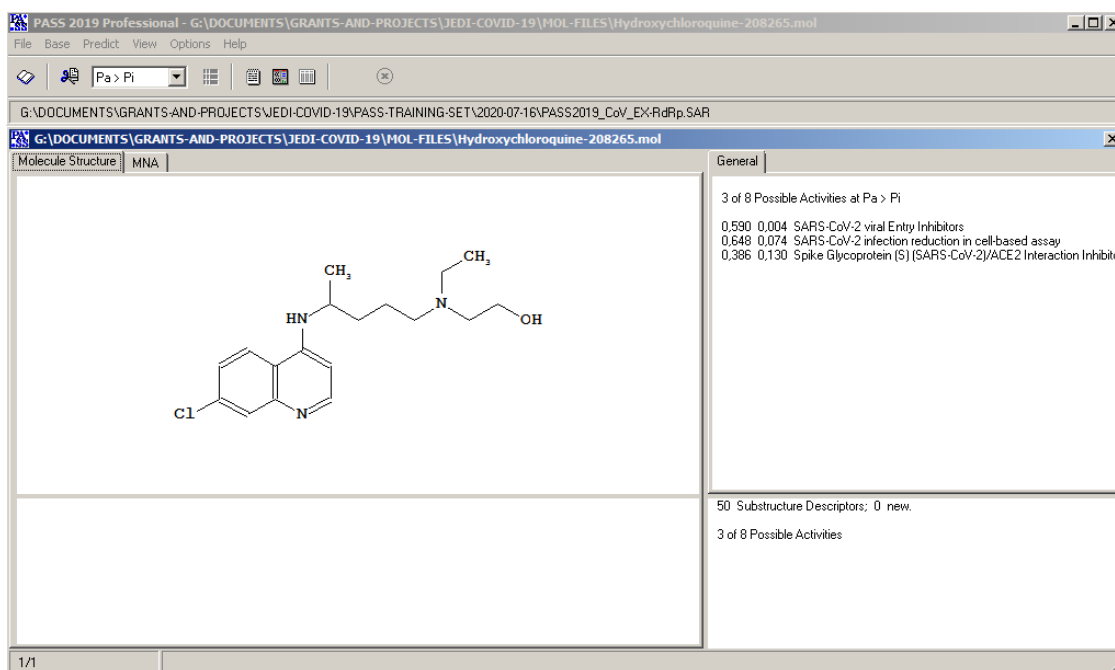
The examples of PASS predictions for the reference substances also demonstrate its reasonable performance:



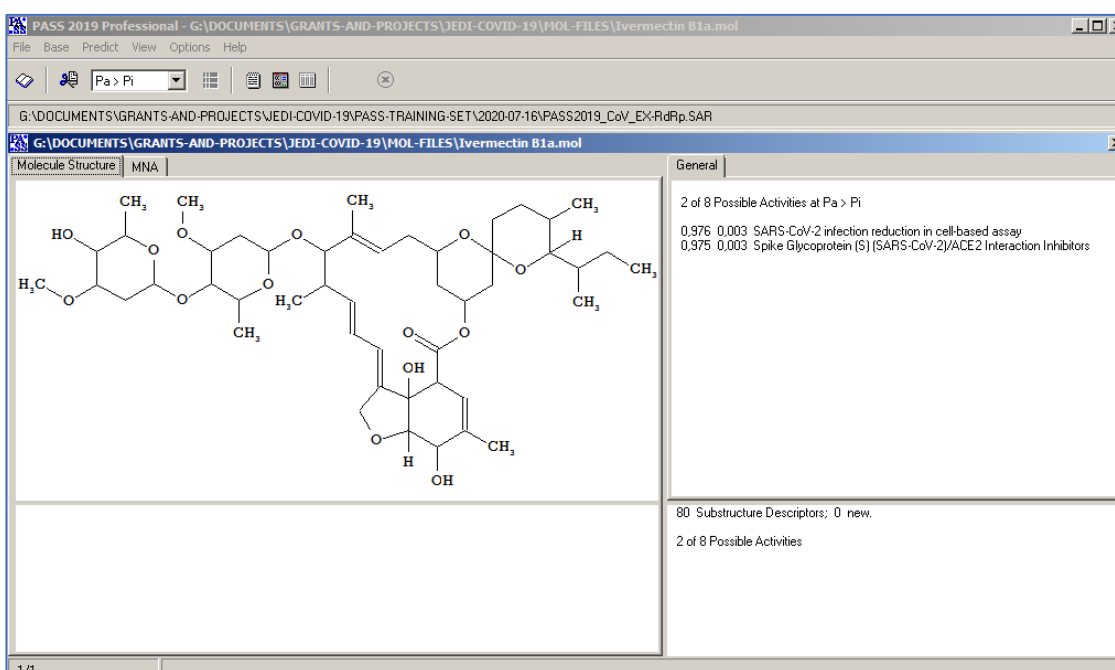
Prediction for Remdesivir



Prediction for Umifenovir



Prediction for Hydroxychloroquine



Prediction for Ivermectin B1a

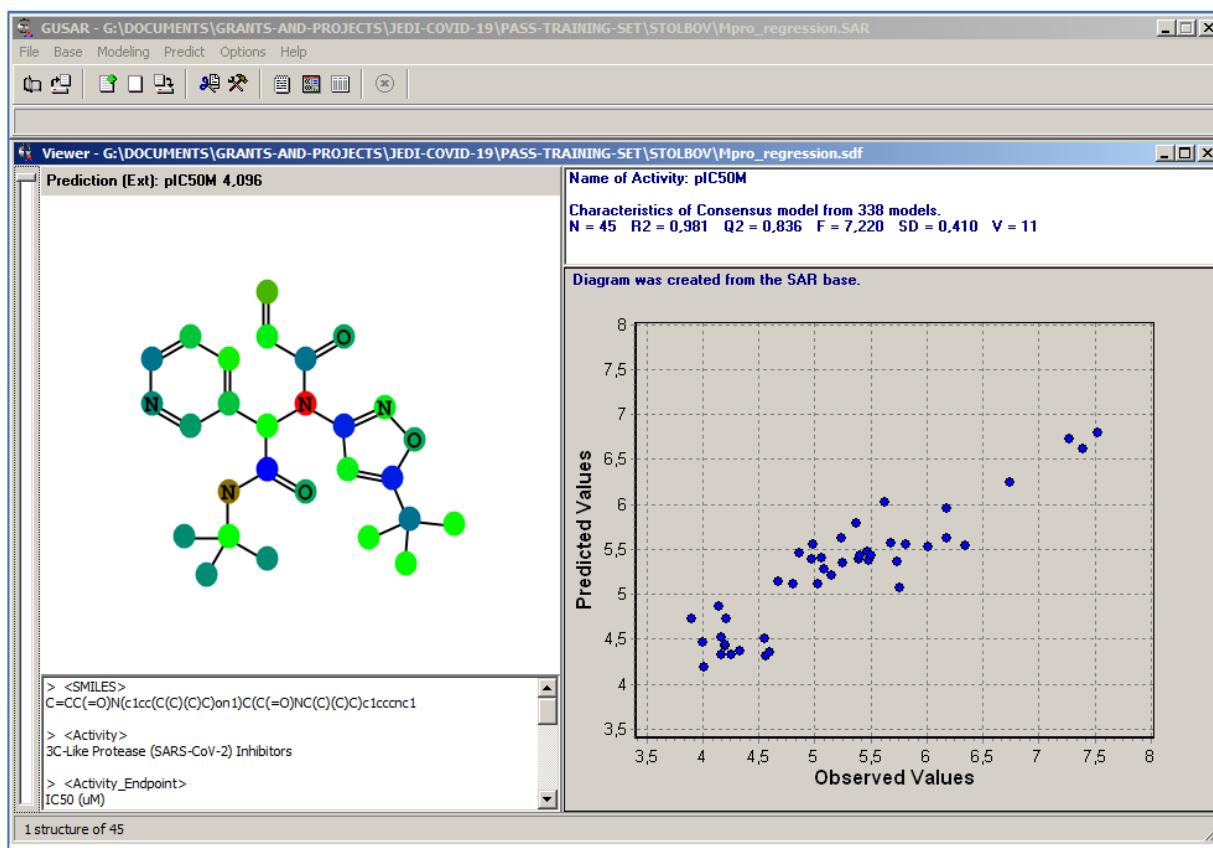
3. Machine learning with computer program GUSAR

GUSAR (General Unrestricted Structure-Activity Relationships) [35] is the software developed for analysis of quantitative structure-activity relationships (QSAR) based on the structural formulas of the compounds and data on their activity/property, and for prediction of activity/property for new compounds. It can be used for the creation of (Q)SAR models for the prediction of properties of organic compounds belonging to both homogeneous and heterogeneous chemical classes. The GUSAR program uses the QNA descriptors that describe the molecule as a set of tuples composed of real values <P,Q> [7]. The P and Q values are calculated for each atom in a molecule under

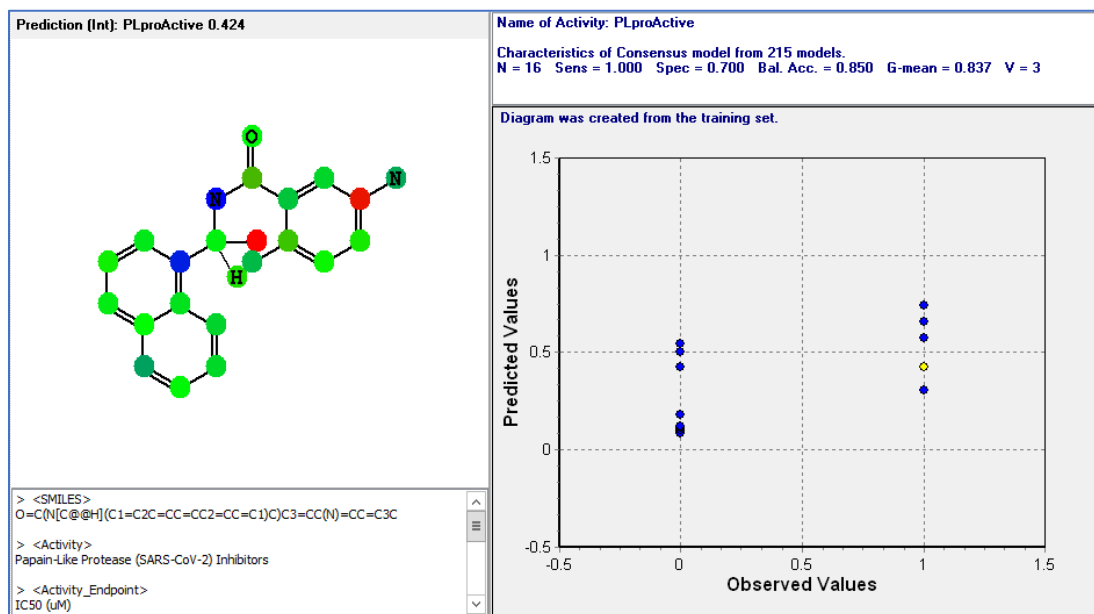
examination using the connectivity matrix and the standard values of the ionization potential of the molecule and the electron affinities of the atoms in the molecule. The current version of the GUSAR program also uses certain physicochemical descriptors and the results of Pa-Pi prediction using the PASS algorithm for 3,663 types of activity and a set including over 300000 biologically active organic compounds. The GUSAR algorithm is based on the self-consistent regression (SCR) method [36]. In the current version of the GUSAR this algorithm is used in combination with the nearest neighbors evaluation and a radial basis function artificial neural network (RBF ANN) based on the SCR results to achieve a multiple-model consensus [37, 38]. A comparative study of the first version of the GUSAR program and the CoMFA, CoMSIA, Golpe/GRID, HQSAR, and other methods that are widely used to construct the QSAR models demonstrated the advantages of the approach used to develop our software [7]. Recently, in the Collaborative Modeling Project for Androgen Receptor Activity (CoMPARA), GUSAR estimations were shown to be quite reasonable [39].

In the JEDI Grand Challenge against COVID-19 using GUSAR, we developed QSAR models for three targets including 3CLpro, PLpro and RdRp.

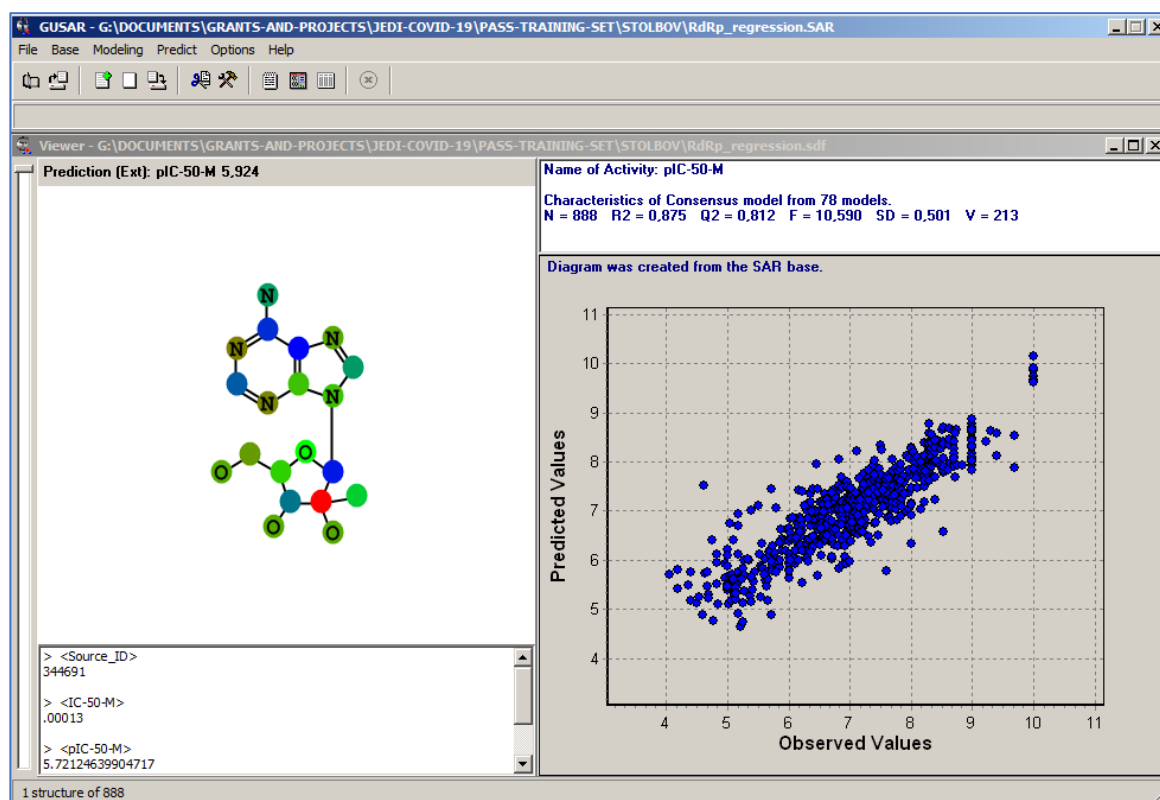
The qualitative characteristics for the 3CLpro QSAR model, which demonstrate the reasonable accuracy and predictive ability of the models, are presented in Figure below.



The qualitative characteristics for the PLpro classification model, which demonstrate the reasonable accuracy and predictive ability of the models, are presented in Figure below.



The qualitative characteristics for the RdRp QSAR model, which demonstrate the reasonable accuracy and predictive ability of the models, are presented in Figure below.



As one may see from the figures presented above, for 3CLpro and RdRp quality of the training set were good enough to create the regression models. In case of PLpro, we were able to develop only classification model with reasonable characteristics of accuracy and predictivity.

4. Molecular modeling for verification of selection

The final verification for the limited number of the selected hits was performed using molecular docking approach for prediction of binding poses and estimation of affinity (using scoring function values). The Docking was performed using Dock 6.5 [40] and AutoDock Vina [41] programs. The cutoff of scoring function for further selection of compounds was chosen -65 kkal/mol and -8.0 kkal/mol for Dock 6.5 and AutoDock Vina, accordingly. The selected binding poses were manually inspected for their ability to accommodate the subpockets in the proteases active sites and analyzed the binding features (H-bonds, steric and electrostatic complementarity).

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 pre-processing (e.g., energy minimization, residue correction, alternative folding, ...) was performed.

We have selected four of six targets proposed by the organizers of the JEDI Grand Challenge against COVID-19 based on the following criteria: (1) vital role of coronavirus entry into the host cell or replication; (2) availability of information about the reference substances for assessing activity by similarity; (3) availability of drug-like compounds data that could be used as the training sets for creating (Q)SAR models; (4) availability of 3D structure in Protein Data Bank. If at least three of the four listed above requirements fulfilled, the target was considered as suitable for the further analysis.

Target 1: 3-chymotrypsin-like protease (3CLpro/Mpro)

The 3CLpro, also known as Nsp5, is first automatically cleaved from polyproteins to produce mature enzymes, and then further cleaves downstream Nsps at 11 sites to release Nsp4-Nsp16. At present, there are many 3D structures of this protease available in PDB. At the first step, all available spatial structures were downloaded from RCSB PDB and analyzed for determination the features participated in binding of inhibitors. For docking approach, the structure **6LU7** with inhibitor N3 was selected as a target. This structure was selected since it contains the largest inhibitor that is similar to the natural substrate. The preparation of protein structure was done using SYBYL 8.1 suite and included: a) deletion of inhibitor, water and cocrystallized ions; b) added hydrogens; c) calculation of atomic charges by Gasteiger-Hückel method; structure optimization by energy minimization in vacuum using Tripos force field.

Target 2: Papain-like proteinase (PLpro)

PLpro is responsible for the cleavages of N-terminus of the replicase polyprotein to release Nsp1, Nsp2 and Nsp3, which is essential for correcting virus replication. The structure **6WUU** was selected as a target for molecular docking. The preparation for docking was the same as for 3CLpro.

Target 3: RNA-dependent RNA polymerase (RdRp)

Nsp12, a conserved protein in coronavirus, is an RNA-dependent RNA polymerase (RdRp) and the vital enzyme of coronavirus replication/transcription complex.

Target 4: Transmembrane peptidase serine 2 (TMPRSS2)

TMPRSS2 cleaves the SARS-CoV-2 spike protein, thus facilitating infectivity of the virus. Unfortunately, no 3D structure of this protein is currently available.

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.

Library 1: ZINC (<https://zinc.docking.org/>) includes 920,839,556 structures. Over 750 million compounds are potentially purchasable.

Library 2: SAVI – Synthetically Accessible Virtual Inventory (https://cactus.nci.nih.gov/download/savi_download/) includes about 1.75 billion proposed products' structures with reactions generated in the first full enumeration of the SAVI project. Number of the synthesizable compounds is about 976 million (621 million without stereoisomers).

Library 3: SWEETLEAD (<https://simtk.org/projects/sweetlead>) includes 9,127 structures (7,636 without stereoisomers).

Library 4: AMS - Aldrich Market Select (<https://www.sigmaaldrich.com/chemistry/chemistry-services/aldrich-market-select.html>) includes 4,787,319 structures, which samples available in stock of Merck KGaA collected in the framework of the program "Antimicrobial Stewardship".

Library 5: Antiviral CAS dataset (<https://www.cas.org/covid-19-antiviral-compounds-dataset>) includes 49,408 structures of antiviral compounds and their analogs collected by Chemical Abstracts Services.

Library 6: Natural Compounds Set includes 118,894 structures of natural compounds collected by our team from several publicly available databases: ChEBI (<https://www.ebi.ac.uk/chebi/>), NNPDB (<http://african-compounds.org/nanpdb/>), NPASS (<http://bidd2.nus.edu.sg/NPASS/>), NuBBE DB (<https://nubbe.iq.unesp.br/portal/nubbedb.html>), UNPD (<http://pkuxxi.pku.edu.cn>).

Library 7: IBS Natural Compounds Set (<https://www.ibscreen.com/>) includes 69,034 structures of natural compounds, their analogs and derivatives, which samples could be purchased from InterBioScreen Ltd.

Library 8: WWAD – World Wide Approved Drugs includes 4,108 structures of the launched drugs prepared by our team in the framework of our project dedicated to drug repurposing.

Library 9: ENAMINE in-stock compounds (<https://enamine.net/>) includes 1.94 million structures that could be obtained from Enamine Ltd.

All data were subjected to the pre-processing and standardization procedures using ChemAxon JChem Instant software and our own computer program ClearSDF in accordance with the contemporary recommendations [42-44].

After the cleaning procedure and selection of the molecules with the higher chance of availability of samples or synthesizability, we obtained the final library with 1,082,000,000 structures that were analyzed by our computational methods described above.

Section 4: results

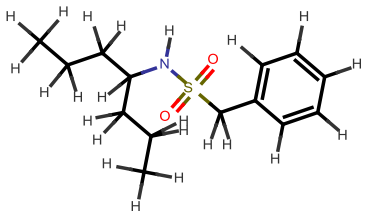
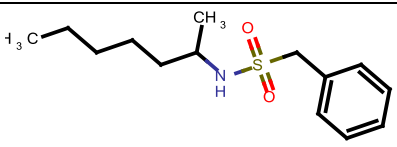
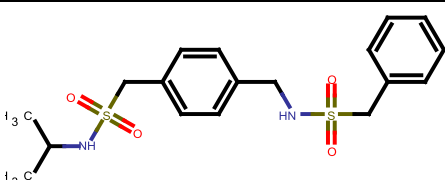
Briefly describe you key findings, any interesting trends in your data, a description of your top 5 compounds for each target. If possible, provide a link to a code and/or data repository. Please do not submit randomly selected compounds!

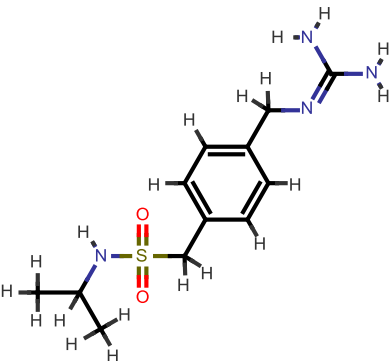
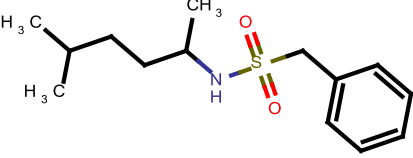
Results: As one may see from Figure 1, based on the assessment of MNA and QNA similarity for the reference molecules described in the Section 1, we selected 42,509 hits, including 12,230 potential 3CLpro inhibitors; 25,812 potential PLpro inhibitors 3,584 potential RdRp inhibitors; and 883 potential TMPRSS2 inhibitors.

Further selection was performed based on the PASS predictions. As a result, we selected 7,148 potential 3CLpro inhibitors; 25,782 potential PLpro inhibitors; 3,544 potential RdRp inhibitors; and 882 potential TMPRSS2 inhibitors.

Since for Transmembrane peptidase serine 2 spatial structure is not available, and for the TMPRSS2 inhibitors we could not create both regression and classification models by GUSAR, this step of the selection was the final.

Top five examples of the selected compounds for **Transmembrane peptidase serine 2 (TMPRSS2)** are given in the table below.

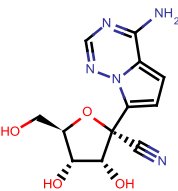
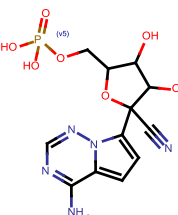
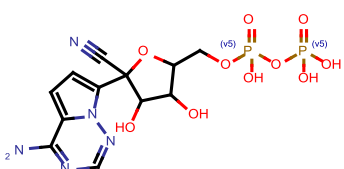
Name (Database)	Structure	Pa-Pi
ZINC001252905755 (ZINC)		0.315
Z355234742 (Enamine)		0.286
Z198103156 (Enamine)		0.280

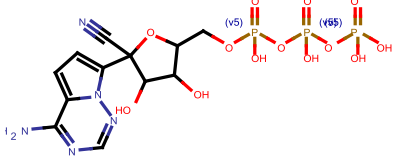
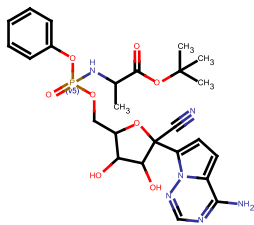
ZINC000261870776 (ZINC)		0.256
Z133631148 (Enamine)		0.250

Since the probability of TMPRSS2 inhibiting activity estimated by PASS is less than 0.4, one may conclude that chances of detecting activity in the experiment are not very high; however, having the prediction confirmed, the identified molecule may become a parent compound of a new chemical class for the studied biological activity (New Chemical Entity) [45].

The further analysis was performed for the other three targets using the regression and classification (Q)SAR models built with GUSAR. As a result, we selected 6,655 potential 3CLpro inhibitors and 3 387 potential RdRp inhibitors with the estimated IC₅₀<10uM. For PLpro, using classification models, we selected 6,981 hits, which according to the prediction are belonging to the class of “actives”.

For RdRp this step of the selection was the final. Top five examples of the selected compounds for **RNA-dependent RNA polymerase (RdRp)** are given in the table below.

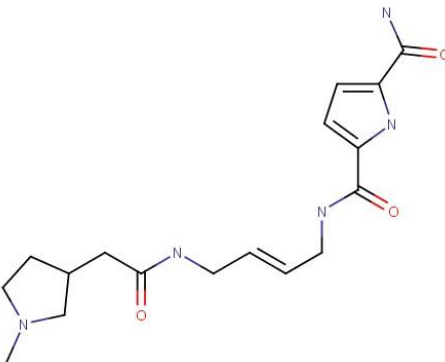
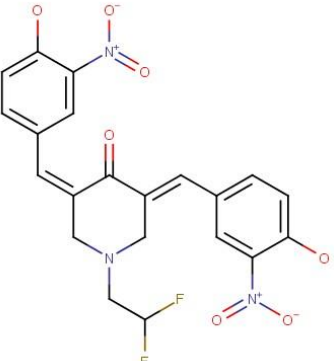
Name (Database)	Structure	Pa-Pi
CHEBI:147281 (Database of natural compounds); BRDWIEOJOWJCLU-LTGWCKQJSA-N (AMS)		0.984
1911578-74-9 (CAS antiviral DB)		0.977
1911578-77-2 (CAS antiviral DB)		0.973

1355149-45-9 (CAS antiviral DB)		0.971
2093124-23-1 (CAS antiviral DB)		0.968

As one may see, Top-5 predicted RdRp inhibitor have very high (Pa-Pi) values; therefore, the chance to confirm this activity in the experiment is significant, but those compounds are very similar to known antiviral drugs (for from the five selected compounds are belonging to the CAS antiviral DB).

For the hits with the potential 3CLpro and PLpro inhibiting activity, additional analysis was carried out using the docking as described above in the Section 1. As a result, we selected 45 potential 3CLpro and 38 potential PLpro inhibitors, for which computational predictions using similarity assessment and machine learning methods implemented in PASS and GUSAR are confirmed by visual inspection.

Top five examples of the selected compounds for **3-chymotrypsin-like protease (3CLpro/Mpro)** are given in the table below.

Name (Database)	Structure	Scoring Function
CHUUJOGSXZEWIU-NSCUHMNNSA-N (AMS)		-66.2 (Dock 6.5)
SPSIFTRUXBQBRF-YOENDLTHSA-N (AMS)		-8.4 (AutoDock Vina)

SXCFTBTXHZXEIN- NRFANRHFSA-N (AMS)	A complex molecule featuring a phthalimide core. One nitrogen is substituted with a 4-(chloromethyl)benzoyl group, and the other is substituted with a 4-(4-chloro-3-isopropoxyphenyl)benzoyl group.	-8.6 (AutoDock Vina)
ZINC001304515852 (ZINC)	A complex molecule with a central 1,2,4-triazole ring. It is substituted with a 4-methyl-1,2,3,4-tetrahydroisoquinoline group, a 4-(trifluoromethyl)pyridin-2-ylmethyl group, and a methanesulfonyl group.	-8.1 (AutoDock Vina)
JNKXJJQRMIPCD- UHFFFAOYSA-N (AMS)	A molecule consisting of a phthalimide ring system connected at the 2-position to a quinoline ring. The quinoline ring has a 2-(chloromethyl)amino group at the 4-position.	-8.4 (AutoDock Vina)

Top five examples of selected compounds for **Papain-like proteinase (PLpro)** are given in the table below.

Name (Database)	Structure	Scoring Function
NIKRPEWINGWQFH- FOWTUZBSSA-N (AMS)	A molecule featuring a 4-methoxyphenyl group connected to a furan ring. The furan ring is substituted with a 2-(2-(4-methylpent-1-en-1-yl)vinyl)amino group.	-8.2 (AutoDock Vina)

DUJJXYLPLPJQH- RVDMUPIBSA-N (AMS)	A chemical structure featuring a naphthalene ring system. A cyano group (-C≡N) is attached to a methine group (=CH-), which is further connected to a thiophene ring. The thiophene ring is substituted with a 4-bromophenyl group.	-9.7 (AutoDock Vina)
ORPOQLQFKDBKIH- UHFFFAOYSA-N (AMS)	A chemical structure consisting of a central amide linkage (-NH-CO-NH-) connecting two side chains. One side chain is a propargyl group (-CH2-C≡CH), and the other is a 4-(trifluoromethyl)phenyl group (-CH2-C6H4-CF3).	-8.2(AutoDock Vina)
RZDABXZTUJAGHM- MDWZMJQESA-N (AMS)	A chemical structure featuring a cyclohexane ring substituted with a carboxylic acid group (-COOH) and a side chain containing an amide linkage (-NH-CO-) connected to a trans-alkene, which is further linked to another amide and a fluorenyl group.	-67.6 (Dock 6.5)
HACZBPORWQBSSJ- OUKQBFOZSA-N (AMS)	A chemical structure featuring a benzimidazole ring system. It is substituted with a sulfonamide group (-SO2NH2) and a side chain containing an amide linkage (-NH-CO-) connected to a trans-alkene, which is further linked to a benzofuran moiety.	-67.8 (Dock 6.5)

Taking into account the requirements of the JEDI COVID-19 Challenge (10,000 hits for each target), to the best scored compounds discussed above, we added more compounds selected with less scores if such data were available.

To estimate the possibility of K_d calculation, we analyzed the availability of K_d data for four targets used for virtual screening of hits that may inhibit SARS-CoV-2 entry in the host cell or viral replication in the Clarivate Analytics CDDI database [34]. No such data were found for 3CLpro, PLpro, RdRp and TMPRSS2 proteins. Thus, it is not possible to estimate the K_d values for the selected hits based on chemical similarity or machine learning approaches.

Despite the recent progress in estimation of ligand-target binding energies by computational methods [46], it is still not accurate enough, particularly for non-congeneric datasets. Calculation of K_d values using docking approaches that we applied in the current study, certainly will not provide the reasonable accuracy [47]. Therefore, we ranked the most promising hits according to the other scores (Pa-Pi values in PASS and IC₅₀ values in GUSAR regression models).

The compounds from our training set for RDRP were analogues of nucleotides. Their mechanism of inhibition of RdRp have proposed incorporation of them in structure of RNA, preventing the further attachment of nucleotides to RNA. The molecular docking programs used in this investigation do not allow predicting the correct poses and estimating binding affinity for inhibitors with such mechanism of action. Thus, the molecular docking approach was not used at the last stage of compounds selection in case of RdRp.

Conclusions

The consequences of the coronavirus infection SARS-CoV-2 / COVID-19 on various aspects of our lives can only be recognized entirely when the pandemic is over. At the same time, the need for a quick response by humanity to a sudden biogenic threat has predetermined the previously unthinkable pace of scientific and clinical research in this area. The virus genome has been decoded, diagnostic tests based on PCR and ELISA have been developed, some mechanisms of pathogenesis have been identified, probable target proteins have been established, experimental models have been created for in vitro testing of potential antiviral compounds, clinical trials are being conducted to reposition drugs, assess the safety and effectiveness of candidate vaccines, and the peculiarities of the patients' response to virus infection and ongoing therapy are studied. A significant part of the results immediately becomes available to the public on the websites of scientific journals and in numerous databases.

Some of the results obtained in the experiment and the clinics are not consistent with each other, which requires further refinement; some of the works published in a fast track manner are not sufficiently substantiated, etc.

For instance, in March 2020 chloroquine/hydroxychloroquine as well as combination of Lopinavir/Ritonavir were proposed as the remedies for treatment of COVID-19. Their anti-SAR-CoV-2 activity was also confirmed in reduction of viral infection in cell-based assays. However, the further clinical trials did not confirm the efficacy of chloroquine/hydroxychloroquine and combination of Lopinavir/Ritonavir; thus, currently World Health Organization does not recommend their use in therapy of SARS-CoV-2 infection.

The uniqueness of the JEDI COVID-19 Challenge is evident because the analysis of experimental and clinical information related to SARS-CoV-2/COVID-19 must be carried out almost simultaneously with its appearance.

Thus, starting from data about dozen compounds that reduce the viral infection in cell assays, later we were able to collect the training sets and to build the classification and regression (Q)SAR models. Those models were applied for virtual screening of hits with potential anticoronaviral activity among 1+ billion drug-like compounds. Information about the selected hits for four anti-SARS-CoV-2 targets is given in four separates *.CSV files.

Experimental testing of the selected hits will significantly increase the knowledgebase in this field, because the current training sets used to develop conventional QSAR models do not exceed thousands of entries, while the estimated size of drug-like chemical space is up to 10^{60} molecules [48]. We expect that with the growth of experimental data and expansion of the studied chemical

space will lead to the discovery of novel medicines with much improved safety and potency profiles for COVID-19 therapy.

Other comments:

Since in the current study we performed virtual screening among 1+ billion molecules from nine databases with very different identifiers, and data regarding the direct links to the vendor in the most cases was not available, we added the field "Vendor" to the *.CSV file of the report.

References

1. Kubinyi H. Chemical similarity and biological activities. *Journal of Brazilian Chemical Society*, 2002, 13 (6), 717-726.
2. Dimova D., Bajorath J. Advances in Activity Cliff Research. *Molecular Informatics*, 2016, 35 (5), 181-191.
3. Wermuth C.G. Similarity in drugs: reflections on analogue design. *Drug Discov. Today*, 2006, 11 (7-8), 348-354.
4. Sheridan R.P., Kearsley S.K. Why do we need so many chemical similarity search methods? *Drug Discov. Today*, 2002, 7 (17), 903-911.
5. Bender A. How similar are those molecules after all? Use two descriptors and you will have three different answers. *Expert Opin. Drug Discov.*, 2010, 5 (12), 1141-1151.
6. Filimonov D., Poroikov V., Borodina Yu., Glorizova T. Chemical similarity assessment through multilevel neighborhoods of atoms: definition and comparison with the other descriptors. *J. Chem. Inf. Comput. Sci.*, 1999, 39 (4), 666-670.
7. Filimonov D.A., Zakharov A.V., Lagunin A.A., Poroikov V.V. QNA based "Star Track" QSAR approach. *SAR and QSAR in Environmental Research*, 2009, 20 (7-8), 679-709.
8. Stolbov L.A., Druzhilovskiy D.S., Filimonov D.A., Nicklaus M.C., Poroikov V.V. (2020). (Q)SAR models of HIV-1 proteins inhibition by drug-like compounds. *Molecules*, 25, 87.
9. Filimonov D.A., Druzhilovskiy D.S., Stolbov L.A., Pogodin P.V., Nicklaus M.C., Poroikov V.V. Assessing biological activity by similarity with MNA and QNA descriptors: Case study of HIV-1 protease, reverse transcriptase and integrase inhibitors. 2020, Unpublished.
10. PostERA activity data URL [https://postera.ai/covid/activity_data]
11. C.Y. Chou, C.H. Chien, Y.S. Han, M.T. Prebanda, H.P. Hsieh, B. Turk, G.G. Chang, X. Chen Thiopurine analogues inhibit papain-like protease of severe acute respiratory syndrome coronavirus, *Biochem. Pharmacol.*, 2008, 75, 1601-1609.
12. Ratia K., Pegan S., Takayama J., Sleeman K., Coughlin M., Baliji S., Chaudhuri R., Fu W., Prabhakar B.S., Johnson M.E., Baker S.C., Ghosh A.K., Mesecar A.D. *PNAS*, 2008, 105 (42), 16119-16124
13. Ghosh A.K., Takayama J., Aubin Y., Ratia K., Chaudhuri R., Baez Y., Sleeman K., Coughlin, M., Nichols, D.B., Mulhearn D.C., Prabhakar B.S., Baker S.C., Johnson M.E., Mesecar A.D. Structure-based design, synthesis, and biological evaluation of a series of novel and reversible inhibitors for the severe acute respiratory syndrome-coronavirus Papain-like protease. *J. Med. Chem.*, 2009, 52(16): 5228.
14. Stanford Coronavirus Antiviral Research Database URL [<https://covdb.stanford.edu/>]
15. Pruijssers A.J., George AS, ... , Sheahan T.P. "Remdesivir potently inhibits SARS-CoV-2 in human lung cells and chimeric SARS-CoV expressing the SARS-CoV-2 RNA polymerase in mice." *bioRxiv*, 2020. doi.org/10.1101/2020.04.27.064279.
16. Bojkova D., McCreig J.E., ... , Cinatl J. "SARS-CoV-2 and SARS-CoV differ in their cell tropism and drug sensitivity profiles." *bioRxiv*, 2020. doi.org/10.1101/2020.04.03.024257.
17. De Meyer S., Bojkova D., ... , Ciesek S. "Lack of Antiviral Activity of Darunavir against SARS-CoV-2." *medRxiv*, 2020. doi.org/10.1101/2020.04.03.20052548.

18. Sheahan T.P., Sims A.C., ... , Baric R.S. "An orally bioavailable broad-spectrum antiviral inhibits SARS-CoV-2 and multiple endemic, epidemic and bat coronavirus." *bioRxiv*, 2020. doi.org/10.1101/2020.03.19.997890.
19. Riva L., Yuan S., ... , Chanda S.K. "A Large-scale Drug Repositioning Survey for SARS-CoV-2 Antivirals." *bioRxiv*, 2020. doi.org/10.1101/2020.04.16.044016.
20. Wang M., Cao R., ... , Xiao G. Remdesivir and chloroquine effectively inhibit the recently emerged novel coronavirus (2019-nCoV) in vitro. *Cell Research*, 2020. doi.org/10.1038/s41422-020-0282-0. [PubMed32020029] [PMC7054408]
21. Ellinger B., Bojkova D., ... , Ciesek S. "Identification of inhibitors of SARS-CoV-2 in-vitro cellular toxicity in human (Caco-2) cells using a large scale drug repurposing collection.", 2020. doi.org/10.21203/rs.3.rs-23951/v1.
22. Sheahan T.P., Sims A.C., ... , and Baric R.S. "An orally bioavailable broad-spectrum antiviral inhibits SARS-CoV-2 and multiple endemic, epidemic and bat coronavirus." *bioRxiv*, 2020. doi.org/10.1101/2020.03.19.997890.
23. ChEMBL database URL [<https://www.ebi.ac.uk/chembl/>]
24. Sielaff F., Böttcher-Friebertshäuser E., Meyer D., Saupe S.M., Volk I.M., Garten W., Steinmetzer, T. Development of substrate analogue inhibitors for the human airway trypsin-like protease HAT. *Bioorganic & Medicinal Chemistry Letters*, 2011, 21(16), 4860-4864.
25. Protein Data Bank URL [<https://www.rcsb.org/>]
26. PASS program package, © Filimonov D.A., Poroikov V.V., Gloziozova T.A., Lagunin A.A. Certificate of Russian State Patent Agency, No. 2006613275 of 15.09.2006.
27. Poroikov V.V., Filimonov D.A., Gloriozova T.A., Lagunin A.A., Druzhilovskiy D.S., Rudik A.V., Stolbov L.A., Dmitriev A.V., Tarasova O.A., Ivanov S.M., Pogodin P.V. Computer-aided prediction of biological activity spectra for organic compounds: the possibilities and limitations. *Russ. Chem. Bull.*, 2019, 68 (12), 2143-2154.
28. Burov Yu.V., Poroikov V.V., Korolchenko L.V. National system for registration and biological testing of chemical compounds: facilities for new drugs search. *Bull. Natl. Center for Biologically Active Compounds (Rus.)*, 1990, No. 1, p.4-25.
29. Filimonov D.A., Druzhilovskiy D.S., Lagunin A.A., Gloriozova T.A., Rudik A.V., Dmitriev A.V., Pogodin P.V., Poroikov V.V. Computer-aided prediction of biological activity spectra for chemical compounds: opportunities and limitations. *Biomedical Chemistry: Research and Methods*, 2018, 1 (1), e00004.
30. Poroikov V.V., Filimonov D.A., Borodina Yu. V., Lagunin A.A., Kos A. Robustness of biological activity spectra predicting by computer program PASS for non-congeneric sets of chemical compounds. *J. Chem. Inform. Comput. Sci.*, 2000, 40 (6), 1349-1355.
31. Geronikaki A., Druzhilovsky D., Zakharov A., Poroikov V. Computer-aided predictions for medicinal chemistry via Internet. *SAR and QSAR in Environ. Res.*, 2008, 19 (1 & 2), 27-38.
32. Anusevicius K., Mickevicius V., Stasevych M., Zvarych V., Komarovska-Porokhnyavets O., Novikov V., Tarasova O., Gloriozova T., Poroikov V. Design, synthesis, in vitro antimicrobial activity evaluation and computational studies of new N-(4-iodophenyl)- β -alanine derivatives. *Research on Chemical Intermediates*, 2015, 41 (10), 7517-7540.
33. Murtazalieva K.A., Druzhilovskiy D.S., Goel R.K., Sastry G.N., Poroikov V.V. How good are publicly available web services that predict bioactivity profiles for drug repurposing? *SAR and QSAR in Environmental Research*, 2017, 28 (10), 843-862.
34. Cortellis Drug Discovery Intelligence URL [<https://www.cortellis.com/drugdiscovery/>]
35. GUSAR (General Unrestricted Structure-Activity Relationships) program package, © Zakharov A.V., Filimonov D.A., Lagunin A.A., Poroikov V.V. Certificate of Russian State Patent Agency, No. 2006613591 of 16.10.2006.

36. Filimonov D.A., Akimov D.V., Poroikov V.V. Method of self-consistent regression in analysis of quantitative structure-property relationships of chemical compounds. *Pharmaceutical Chemistry Journal*, 2004, 38 (1) 21-24.
37. Zakharov A.V., Peach M.L., Sitzmann M., Nicklaus M.C. A new approach to radial basis function approximation and its application to QSAR. *J. Chem. Inf. Model.*, 2014, 54 (3), 713-719.
38. Lagunin A.A., Romanova M.A., Zadorozhny A.D., Kurilenko N.S., Shilov B.V., Pogodin P.V., Ivanov S.M., Filimonov D.A., Poroikov V.V. Comparison of Quantitative and Qualitative (Q)SAR Models Created for the Prediction of Ki and IC50 Values of Antitarget Inhibitors. *Frontiers in Pharmacology*, 2018, 9, 1138.
39. Mansouri K., Kleinstreuer N., Abdelaziz A.M., Alberga D., Alves V.M., Andersson P.L., Andrade C.H., Bai F., Balabin I., Ballabio D., Benfenati E., Bhattacharai B., Boyer S., Chen J., Consonni V., Farag S., Fourches D., García-Sosa A.T., Gramatica P., Grisoni F., Grulke C.M., Hong H., Horvath D., Hu X., Huang R., Jeliaskova N., Li J., Li X., Liu H., Manganelli S., Mangiatordi G.F., Maran U., Marcou G., Martin T., Muratov E., Nguyen D.-T., Nicolotti O., Nikolov N.G., Norinder U., Papa E., Petitjean M., Piir G., Pogodin P., Poroikov V., Qiao X., Richard A.M., Roncaglioni A., Ruiz P., Rupakheti C., Sakkiah S., Sangion A., Schramm K.-W., Selvaraj C., Shah I., Sild S., Sun L., Taboureau O., Tang Y., Tetko I.V., Todeschini R., Tong W., Trisciuzzi D., Tropsha A., Van Den Driessche G., Varnek A., Wang Z., Wedebye E.B., Williams A.J., Xie H., Zakharov A.V., Zheng Z., Judson R.S. CoMPARA: Collaborative Modeling Project for Androgen Receptor Activity, *Environmental Health Perspectives*, 2020, 128 (2), 027002-1 - 027002-17.
40. UCSF Dock URL [<http://dock.compbio.ucsf.edu/>]
41. AutoDock Vina URL [<http://vina.scripps.edu/>]
42. Fourches D., Muratov E., Tropsha A. Trust, but verify: on the importance of chemical structure curation in cheminformatics and QSAR modeling research. *J. Chem. Inf. Model.*, 2010, 50(7), 1189-1204.
43. Fourches D., Muratov E., Tropsha A. Curation of chemogenomics data. *Nat. Chem. Biol.*, 2015, 11(8), 535.
44. Fourches D., Muratov E., Tropsha A. Trust, but Verify II: A Practical Guide to Chemogenomics Data Curation. *J. Chem. Inf. Model.*, 2016, 56(7), 1243-1252.
45. Filimonov D.A., Lagunin A.A., Glorizova T.A., Rudik A.V., Druzhilovskii D.S., Pogodin P.V., Poroikov V.V. (2014). Prediction of the biological activity spectra of organic compounds using the PASS online web resource. *Chemistry of Heterocyclic Compounds*, **50** (3), 444-457.
46. Cournia Z., Allen B., Sherman W. Relative Binding Free Energy Calculations in Drug Discovery: Recent Advances and Practical Considerations. *J. Chem. Inf. Model.*, 2017, 57(12), 2911-2937.
47. Chen Y.C. Beware of docking! [published correction appears in Trends Pharmacol. Sci. 2015 Sep;36(9):617]. *Trends Pharmacol. Sci.*, 2015, 36(2), 78-95.
48. Muratov E.N., Bajorath J., Sheridan R.P., Tetko I., Filimonov D., Poroikov V., Oprea T., Baskin I.I., Varnek A., Roitberg A., Isayev O., Curtalolo S., Fourches D., Cohen Y., Aspuru-Guzik A., Winkler D.A., Agrafiotis D., Cherkasov A., Tropsha A. QSAR Without Borders. *Chemical Society Reviews*, 2020, Advance Article. DOI: 10.1039/d0cs00098a