HHS Public Access

Author manuscript

Antiviral Res. Author manuscript; available in PMC 2024 September 01.

Published in final edited form as:

Antiviral Res. 2023 September ; 217: 105620. doi:10.1016/j.antiviral.2023.105620.

Small Molecule Antiviral Compound Collection (SMACC): A comprehensive, highly curated database to support the discovery of broad-spectrum antiviral drug molecules

Holli-Joi Martin[1], **Cleber C. Melo-Filho**[1], **Daniel Korn**[1], **Richard T. Eastman**[2], **Ganesha Rai**[2], **Anton Simeonov**[2], **Alexey V. Zakharov**[2], **Eugene Muratov**[1], **Alexander Tropsha**[1]

[1]UNC Eshelman School of Pharmacy, University of North Carolina, Chapel Hill, NC, USA.

[2]Division of Preclinical Innovation, National Center for Advancing Translational Sciences, Rockville, MD, 20850

Abstract

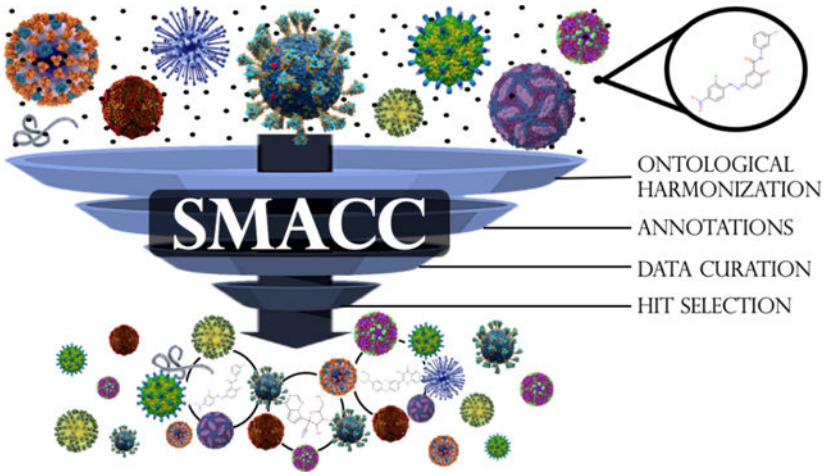
Diseases caused by new viruses cost thousands if not millions of human lives and trillions of dollars. We have identified, collected, curated, and integrated all chemogenomics data from ChEMBL for 13 emerging viruses that hold the greatest potential threat to global human health. By identifying and solving several challenges related to data annotation accuracy, we developed a highly curated and thoroughly annotated database of compounds tested in both phenotypic and target-based assays for these viruses that we dubbed SMACC (Small Molecule Antiviral Compound Collection). The pilot version of the SMACC database contains over 32,500 entries for 13 viruses. By analyzing data in SMACC, we have identified ~50 compounds with polyviral inhibition profile, mostly covering flavi- and coronaviruses. The SMACC database may serve as a reference for virologists and medicinal chemists working on the development of novel BSA agents in preparation for future viral outbreaks.

Graphical Abstract

alexey.zakharov@nih.gov .

6.Conflict of interest

AT and ENM are co-founders of Predictive, LLC, which develops computational methodologies and software for toxicity prediction. All other authors declare they have nothing to disclose.



Keywords

Broad-spectrum; Antiviral Drug Discovery; Database; Drug Repurposing

1. Introduction

Broad-spectrum antiviral (BSA) drugs could protect against emergent viruses; however, the development of such drugs has been challenging. As of today, there are only 90 approved antiviral drugs of which only 11 are approved to treat more than one virus (Erik and Guangdi, 2016). Most approved antiviral drugs are effective against herpes, hepatitis, or human immunodeficiency viruses, but offer no protection against the recent SARS-CoV-2 pandemic. However, the fact that medications active against more than one virus do exist fuels the expectation that such medications can be developed in principle via concerted strategic effort.

Despite the clear need for BSA medications, previous outbreaks have shown that the interest in supporting viral research and drug discovery vanishes quickly after about a year past the viral threat, leaving the work toward an effective medication unfinished (Bobrowski et al., 2020). To support the focused development of BSAs, we endeavored to collect, curate, and integrate all publicly accessible data on compounds tested in both phenotypic and target-based assays for emerging viruses of concern. To this end, we have (i) conducted a comprehensive evaluation of viruses holding the greatest potential threat to global human health, (ii) used the data available in ChEMBL, an online collection of bioactive molecules with drug-like properties (Gaulton et al., 2017), to build a curated, annotated, and publicly available database of compounds tested in both phenotypic and target-based assays for these viruses, and (iii) identified compounds with BSA activity. We dubbed this pilot-database Small Molecule Antiviral Compound Collection (SMACC) and made it publicly available online at https://smacc.mml.unc.edu. We expect that SMACC database can support further computational and experimental medicinal chemistry studies targeting rational design and discovery of novel BSAs.

2. Methods

2.1. Selection of viruses of interest and initial database generation

Viruses with a high replication rate, especially coupled with genetic mutability and segmented RNA viruses permit the virus to evolve rapidly and gain attributes, including increased transmission or evading preexisting immunity also facilitating outbreak or pandemic spread. The Therefore, we selected the following 13 viruses from five viral families: *Coronaviridae* (SARS-CoV-2, MERS-CoV, HCoV-229E), *Orthomyxoviridae* (H1N2, H7N7), *Paramyxoviridae* (RSV, HPIV-3), *Phenuiviridae* (Sandfly Fever), and *Flaviviridae* (Dengue, Zika, Yellow Fever, Powassan, West Nile) (Gaulton et al., 2017). The viral families under study were selected due to their potential for future viral emergence events (Sessions et al., 2022) whereas the viruses selected as representatives of these families were selected based on the availability of experimental assays capable of validating our hypotheses.

2.2 Data Curation

All data were extracted from ChEMBL 29 (Gaulton et al., 2017). The virus name, and any known alias were used as keywords to extract all phenotypic and target-based assays and compounds tested in these assays for each virus. For the target-based assays, we ran an additional search using virus and target name as the keywords to ensure no respective viral data was lost. To identify drug targets for each virus we searched existing literature using the keywords " [virus_name] virus drug targets". After extraction, the data for each virus were pre-processed and curated as described below. We followed well-established protocols for chemical and biological data curation described by Fourches et al. (Fourches et al., 2016, 2015, 2010). When examining the resulting datasets, we have identified a need for additional curation of assay annotations as discussed in the Results section. We summarized our database entries after curation in Table S1.

The final spreadsheet reports assay results and respective activity call (i.e., active, inactive, or inconclusive) per compound, for each phenotypic and biochemical assay we collected from ChEMBL and analyzed. The spreadsheet also contains full activity profiles based on "final activity calls". Detailed rules for making the final activity calls for compounds tested in multiple assays can be found in the supporting information (Table S2). One may see that conflicting activity calls for the same compound is not uncommon. This is explained by high variability of compound testing results even in similar assays, which may be caused by (even small) changes in experimental conditions.

3. Results

3.1 Ontological examination and curation of assays reported in ChEMBL

3.1.1 Phenotypic assays.—While ChEMBL does an exceptional job at providing the largest curated and publicly available bioactivity database, we have identified multiple issues requiring additional data curation efforts to yield a clean database of antiviral activity data. One major challenge was associated with the use of BioAssay Ontology annotations " ("The BioAssay Ontology (BAO)," 2022) for the assay type. For 9 of 13 viruses the BAO assay

type was recorded in ChEMBL as “Organism-Based”. Closer inspection revealed that the assays were run in the whole cell (i.e., “Cell-Based”) format and the outcome was recorded as inhibition of the “organism” (i.e., virus). This highlights the importance of careful data processing both by chemical bioactivity data curators, as well as by the user looking to extract data from large data collections. For example, if one were to search ChEMBL for “cell-based assays” for these viruses, 99.44% (27,410 of 27,562 entries) of the data would not be identified. These semantic ambiguities amongst various annotations presents major challenges for users searching for relevant data.

We procured and enriched the original data annotations found in ChEMBL (Table S3), which greatly increased the quality and usability of the extracted data. This step was absolutely necessary to improve both the accuracy of chemical structures (which ChEMBL has substantially improved over the years) (“ChEMBL Compound Curation Pipeline,” 2020; Papadatos et al., 2015) and correctness of activity labeling such that users can obtain the entirety of existing but effectively, hidden data which they searched for.

Missing, i.e., absent from their designated entry field, data annotations were also extremely common. Despite there being a distinct field for the respective entry, 13.72% of all phenotypic assays results did not indicate which cell type was used. Instead, the cell type was found in the assay descriptions, where we could extract and properly annotate the field. However, 36.73% of all missing cell types were not listed in the assay description either. This required manually searching the literature for the cell type used, which was extremely time consuming, and, in some cases, still, no clear cell type could be identified. If the cell type was not identified, it was annotated as “unclear”. These cases are reported in Table S3 as “Cell Type Completely Missing”. This tedious work resulted in the additional recovery of ~4% of all phenotypic assay results. Another issue of missing data annotations was uncovered when we looked into the class of assays. Most assays were not labeled to indicate whether they were primary, counter, or cytotoxicity assays. Furthermore, the assay descriptions also failed to provide an appropriate level of detail. Many assay descriptions simply reported “Antiviral activity against virus X”. Lacking assay details makes it impossible to analyze data reproducibility and prohibits meaningful integration of multiple assay results.

3.1.2 Target based assays.—

While many of the issues discussed above for phenotypic assays were not present in the target-based assays, some cases needed further attention. One example includes 536 entries deposited as compounds tested against “genome polyprotein” of West Nile or Zika viruses. However, upon closer examination of the ChEMBL records, we have established that these compounds were actually tested against the NS2B-NS3 protease, rather than the entire genome polyprotein.

3.2 Development of the curated data entries in the SMACC database

Our extensive curation efforts following the protocols described in Methods led to the removal of compounds from each viral family in the phenotypic data set (Figure 1). From the initial compound list through the normalization of specific chemotypes, we removed ~26 % of compounds (n=4801) from the *Coronaviridae* family, 25% (n=6) from the

Phenuiviridae, ~21% (n=594) from *Flaviviridae* and ~8% (n=7) from *Orthomyxoviridae* and ~8% (n=350) from *Paramyxoviridae*.

For the target-based curation, there were fewer compounds removed due to inconsistent data and molecular cleaning (Figure 2). Here the family where the largest number of compounds were removed was also the *Coronaviridae* which was less than 1% of all compounds (n=13). Interestingly, our annotation efforts followed a similar pattern, where the target-based data were deposited with more annotations, therefore requiring less curation than the phenotypic data.

Note that at this step of data curation we intentionally kept duplicative compound records reflecting our objective to check whether the same compound showed similar activity against different viruses (i.e., had a potential to be a broad-spectrum agent). However, such chemically duplicative entries have been annotated in SMACC to facilitate their removal prior to the development of assay specific QSAR models by the users of SMACC.

3.3 Curated Phenotypic Data

Curated phenotypic testing entries in our database included assay data for 13 viruses in 5 viral families: *Coronaviridae* (SARS-CoV-2, MERS-CoV, HCoV-229E), *Orthomyxoviridae* (H1N2, H7N7), *Paramyxoviridae* (RSV, HPIV-3), *Phenuiviridae* (Sandfly Fever), and *Flaviviridae* (Dengue, Zika, Yellow Fever, Powassan, West Nile).

3.3.1 Distribution of compound activity—The heatmap presented in Figure 3 depicts the activity spectrum of all compounds tested in phenotypic assays for the viruses in our database. It is evident that the majority of compounds were either inactive (80.6%) or untested. In contrast, the number of actives constituted 15.6% of the total number of entries, and the fraction of “true actives” with no conflicting assay results was just 6.48% (1,387 compounds). Unsurprisingly, the virus with the largest number of tested compounds was SARS-CoV-2, encompassing 61.6% of our phenotypic assay entries; however, 94.76% of compounds were reported as inactive. Each virus had the prevalence of inactive compounds in the dataset except Dengue (Figure 3). Distribution of compounds by assay types is discussed in Supplemental Information and summarized in Figure S1 and Table S4. We note that the annotation of a compound as active or inactive against any virus should be always reported strictly in the context of the specific underlying assay.

3.3.2 Identifying compounds active in multiple assays.—The full details of how compound activity was annotated is provided in the supporting information (Table S2). We first selected a subset of compounds from our database that were tested in assays against two or more viruses. A matrix was created where each compound occupied one row, and the columns contained concatenated lists of every virus the compound was reported active or inactive against. We systematically analyzed this matrix to identify compounds that we considered to be true actives, i.e., the compound was reported active in all assays it was tested in. Effects of cell types on activity were also analyzed (Figure S1; Table S4).

We report the eight compounds with the highest numbers of activity profiles resulting from the initial analysis in Table 1. Our top compound is CHEMBL4437334 with activity against

Dengue, West Nile, Yellow Fever, and Zika. It is a research compound not yet progressed into any clinical trial, unlike many compounds on this list including CHEMBL4454780 (active against RSV, MERS-CoV, Dengue, and Zika), CHEMBL2016757 (active against RSV, HPIV-3, and Dengue), CHEMBL4544911 and CHEMBL4562509 (active against Dengue, West Nile, and Yellow Fever). Three named compounds were identified from our search: 6-azauridine (active against RSV, West Nile, Dengue); amodiaquine (active against SARS-CoV-2, Dengue, Zika), which is an approved drug for malaria; and brequinar (active against Dengue, West Nile, and Yellow Fever), which is currently in Phase I clinical trials for the treatment of acute myeloid leukemia (Andersen et al., 2019; Li et al., 2019; Park et al., 2020).

Interestingly, brequinar also recently underwent phase II clinical trials against SARS-CoV-2 ("CRISIS2: A Phase 2 Study of the Safety and Antiviral Activity of Brequinar in Non-hospitalized Pts With COVID-19 – Study Results – ClinicalTrials.gov," 2022). While the clinical trial was not successful using brequinar alone, this drug was found highly effective in combination with remdesivir or molnupiravir (Schultz et al., 2022). Research suggests that brequinar's antiviral activity against SARS-CoV-2 is through inhibition of the host cell dihydroorotate dehydrogenase (DHODH) rather than being a direct-acting antiviral. The combination of a nucleobase antiviral and DHODH, or other compounds that impact *de novo* nucleotide synthesis would, in effect, increase the nucleobase antiviral cellular concentration thereby increasing the rate of incorporation into the viral synthesized RNA, in theory. This approach is effective against multiple viruses *in vitro*, for example, brequinar has been shown to inhibit dengue, enterovirus, and Ebola viruses through this same, host-targeted mechanism (Fu et al., 2020; Luthra et al., 2018; Wang et al., 2011). Given the reported activity of brequinar against three *flaviviruses* (Dengue, West Nile, Yellow Fever) in phenotypic assays, we hypothesize the assays were detecting the human dihydroorotate dehydrogenase inhibition, rather than inhibition of a viral target. As such, a future release of SMACC will include an analysis of all phenotypic assays, host-target assays and the overlap between the phenotypic assays and the host-target assays. It is our hope this analysis will help identify potential host-targeting broad-spectrum antiviral drugs.

One of the reviewers of our manuscript pointed out that DHODH inhibitors have been shown in one case to have activity in cell-culture but lose activity *in vivo* due to the use of nucleotide salvage pathways that obviate the need for DHODH during virus replication. (Wang et al., 2011). However, it is important to remember that SMACC has been designed as medicinal chemistry oriented database for early stage drug discovery for antiviral research and only accounts for compound activity in phenotypic and target-based assays. Clinical data and *in vivo* assays are not included in this pilot version of the database; however, such assays may indeed provide greater insights into the activity or mechanism of action for hit compounds selected from SMACC. Thus, detailed interpretation and understanding of the mechanisms of action of the selected compound(s) is left at the discretion of the user. Users are highly encouraged to research in vivo and clinical information on hit compounds to ensure usability for their specific applications prior to experimental testing. These later stages of drug optimization should arise after the compound has confirmed activity in a user's specific phenotypic or target-based assay.

There were 55 additional compounds active against at least two viruses (Table S5). Of these, seven compounds were active against different viral families: three were active against *Paramyxoviridae* (RSV, HPIV-3); two were active against *Coronaviridae* (MERS-CoV, HCoV-229E); and 43 against any two of our *Flaviviridae* viruses (Dengue, Zika, Yellow Fever, Powassan, and West Nile). We also identified 1,324 compounds active against one virus. We also performed a clustering analysis to identify untested compounds with similar structures to compounds with multi-viral activity (Figure S2, Figure S3, Table S6).

3.4 Target-Based Data

Curated target-based testing entries (11,123) in our database include assay data for ten viruses in five viral families: *Coronaviridae* (SARS-CoV-2, MERS-CoV, HCoV-229E), *Orthomyxoviridae* (H7N7), *Paramyxoviridae* (RSV, HPIV-3), *Phenuiviridae* (Sin Nombre), and *Flaviviridae* (Dengue, Zika, West Nile).

3.4.1 Activity of Compounds—Based on the activity calls as reported in the assay results (cf. Methods), most compounds (89.8%) were inactive (Figure 4). Many compounds inactive against SARS-CoV-2 were tested because of the multiple recent campaigns including drug repurposing screenings due to the current pandemic. While SARS-CoV-2 was the most tested virus, three flaviviruses were also well represented in the database (Dengue, West Nile, and Zika); however, as the number of compounds tested increased there was a decrease in the fraction of the active compounds for these viruses. Overall, active compounds represented only ~9.9% of our total dataset where 5.78% (644 compounds) were “true actives”.

3.4.2 Analysis of Targets—The Main Protease ($3CL^{pro}$) of *Coronaviridae* was, unsurprisingly, the most studied target (78.8% of the entries), followed by NS2B-NS3 Protease of *Flaviviridae* (16.2%), NS5 of *Flaviviridae* (1.26%), Integrin alpha-V/beta-3 of *Hantaviridae* (1.17%), and Fusion glycoprotein F0 of *Paramyxoviridae* (1.1%). Interestingly, the virus with the greatest number of targets tested (five) was MERS-CoV and was tested against the spike protein, RDRP, Nucleocapsid protein, M^{pro}, and PL^{pro}.

3.4.3 Analysis of Compounds—We followed the same approach for analyzing the target-based dataset for BSA activity, as was taken for the phenotypic dataset. In this case, the intermediate table included a row for each compound, and the columns were concatenated lists of every virus and target the compound was reported active or inactive against. Our analysis identified 16 compounds active against two viruses at the protein target level (Table 2). Two of these compounds (CHEMBL4544781 and CHEMBL4522602) were active against targets from two different viral families (Zika’s NS5 and MERS-CoV’s RDRP), whereas the others were active against two flaviviruses NS2B-NS3 Protease. We also identified 628 compounds active against one virus and performed structural *clustering of all compounds* (Figure S4, Figure S5, Table S7).

3.5 Concordance between the phenotypic and target-based data

We analyzed the concordance between the 5,934 compounds tested in both phenotypic and target-based assays to: (i) expand our list of hits by identifying potentially promising

compounds that may not have been tested yet, and (ii) hypothesize the mechanism of actions of compounds active both in a virus in a live cell and in a complementary viral target. Our analysis indicated that 35 compounds were active in at least one phenotypic and one target-based assay (Table S8). In many cases, the active calls were within the same viral family. For example, CHEMBL4522006 was active against Dengue Virus in a phenotypic assay, and active against the Dengue NS2B-NS3 protease in a target-based assay. Our data strongly supports the hypothesis that CHEMBL4522006 is active against Dengue virus in the live cell assay by inhibiting its NS2B-NS3 protease, which supports the use of the protease assays for future experimental and computational structure-activity relationship studies. Promising potential BSA compounds, including CHEMBL4522006, are summarized in Table S9.

In other cases, as we observed for CHEMBL4437334, a compound was active against several viruses of the same family in phenotypic assays (Dengue Virus, West Nile Virus, Yellow Fever Virus, Zika Virus) and only tested and active against a subset of those viruses in the target-based assays (NS2B-NS3 Protease of Dengue and Zika). In these cases, we could suggest the compound be tested against the same target in the untested yet highly homologous viruses from the same family, using the principle of viral protein conservation (Melo-Filho et al., 2022).

There were also instances of compounds, such as CHEMBL267099 (tubercidin), reported active in a phenotypic assay for a virus in one family (HPIV-3) and active in a target-based assay of another (Zika NS5 protein). Cases like these are particularly interesting, because after making this connection one can suggest testing this compound in various HPIV-3 targets, live cell Zika virus, as well as the other highly homologous *Flaviviridae* members (West Nile, Yellow Fever, Dengue) in live cell assay and against the NS5 protein.

Our concordance analysis also revealed 52 compounds active in at least one phenotypic assay and inactive in a (supposedly, relevant) target-based assay (Table S10, Supplementary Material). In this case, we recommend compounds be tested in additional target-based assays of the same viral family; it is also possible that the activity of such compounds inactive in viral targeting assays but active in phenotypic assays is actually due to their host-directed mechanism of action. There were also 191 compounds inactive in a phenotypic assay and active in at least one target-based assay (Table S11, Supplementary Information). This could be due to an issue with permeability, a disconnect between biochemical and cell-based assays, or could be because the assays have a different read out. Mapping the virus and viral family of the phenotypic result to the active result of the target helped identify potential new viral families for phenotypic testing, as well as highlighted the importance of whether the cell type used in the phenotypic assay appropriately represented the virus and the antiviral result. Of course, due to the proportion of inactive compounds in our dataset, most compounds (5,656) were concordantly reported as inactive in both phenotypic and target-based assays.

3.6 Integrated, searchable SMACC Database

Our pilot-SMACC database is currently available at https://smacc.mml.unc.edu. Users will find freely downloadable excel sheets containing our phenotypic, target-based, and overlapping datasets including tabs containing subsets of active compounds selected from

the approach described in Methods (see detailed description of the spreadsheets in the Supplemental Information). These excel sheets were designed so that users could easily extract subsets of the database using various filtering options. These filters include molecule (ChEMBL ID, smiles, InChiKey), virus, cell or target type, activity (activity call, raw assay result), and assay type.

The extraction approach described in Methods, i.e., removing compounds with conflicting activity calls from our final BSA activity analysis, was stringent and resulted in a concise list of potential BSA compounds that we had the highest confidence in. We acknowledge the widely varied objectives across antiviral research and emphasize the versatility of this database. Important subsets of BSA compounds identified using different criteria can be found in Supplemental Information (Tables S12, S13). However, we should stress that SMACC has been designed as medicinal chemistry-oriented database for early-stage drug discovery for antiviral research and only includes data for compound activity in phenotypic and target-based assays. Users may identify *in vivo* and clinical information which may provide greater insights into the activity or mechanism of action for hit compounds selected from SMACC. Furthermore, there is always a delay between publishing new data on compound testing in various assays and incorporation of this data into ChEMBL and, consequently, SMACC; thus, we expect users to identify data reported, at any point, in most recent literature but not (yet) included in SMACC.

In addition, we identified 53 of 90 approved drugs in our phenotypic dataset and 57 of the 90 drugs in our target-based dataset (Table S14). Drugs with reported active assay results are summarized in Table 3. Clearly, these drugs have broader activity spectrum than what they are approved for. Further experimental testing based on hypotheses from this table will be extremely valuable to understanding their broad-spectrum potential.

Following this release of our pilot version of SMACC database, we plan to develop and deliver an expanded version of SMACC next year. The SMACC database will undergo a transformation to enable better searchability, filtering, and data exportation. This will include the ability to search and filter by chemical, chemical similarity, assay, cell type, target-type, and more. We will also enable online search and filtering criteria to allow users the ability to isolate and export relevant subsets of the data. More importantly, our content will be updated and expanded upon following each ChEMBL release. We will incorporate PubChem assay data for our current set of viruses, ensuring consistency in data curation, annotation, and proper ontological harmonization with the framework we are using. We also plan to launch compound collections for promising viral targets involved in viral replication such as helicases, methyltransferases, polymerases, and proteases.

4. Discussion

We have collected, curated, and integrated all the chemogenomic data available for a subset of viruses of interest in ChEMBL to identify BSA compounds. We created a pilot version of the SMACC database based on phenotypic and target-based assays in ChEMBL. It adds to a variety of other important datasets and databases listed in Supplemental Information (Table S15). These data collections along with several research initiatives such as the Antiviral

Program for Pandemics (https://www.niaid.nih.gov/research/antivirals) and the Rapidly Emerging Antiviral Drug Development Initiative (READDI; https://www.readdi.org) push the scientific community to work in an 'open science' format.

Our work can guide future data collection to increase the clarity and accessibility of relevant information to a broader scientific community including additional data on other emerging viruses. As antiviral assay data continues to be published online, we must improve the accuracy of assay reporting. This is not only a task for data scientists, but also the broader antiviral community. Antiviral assays are complex, but poorly described in online collections, due to the difficulty in extracting relevant assay information from publications. Descriptions of antiviral assay's reported in literature should be clearly defined and obviously linked to the associated data set so that it can be accurately and easily added to online repositories. Antiviral data ontology and annotation frameworks need to be as descriptive as possible. Most importantly, using a single information type per column, and breaking complex data into different columns as needed. For example, instead of generalizing to "genome polyprotein", the target should be defined as the protein of study in the assay (e.g., NSP13 Helicase). Taking these steps can help improve data quality and accessibility, increasing the benefit to a wide variety of researchers.

The SMACC database also has the potential to guide more informed drug repurposing efforts, which was a popular strategy employed during the first year of the SARS-CoV-2 pandemic. Repurposing FDA approved drugs (Alves et al., 2021) and their combinations (Bobrowski et al., 2021) quickly provided options for clinical use without the need to undergo extensive toxicological testing. For instance, we have identified anticancer drug brequinar as a potential antiviral agent (cf. Table 1), and the analysis of additional bioactivities, including those against host targets, may reveal novel interesting compounds.

Beyond drug repurposing, another intuitive approach that proved useful in the current SARS-CoV-2 pandemic was to identify BSA drugs through proteome conservation analysis, as reported by Schapira et al. (Yazdani et al., 2021) as well as by our group (Melo-Filho et al., 2022). We feel exploring the conservation between homologous coronaviral proteins is an extremely valuable strategy for target selection that could assist the development of BSA compounds.

With viral protein conservation as a tool for identifying BSAs, one wonders if there may be a link between protein conservation and ligand promiscuity. While in theory, the framework of our database would easily allow for this analysis, the unfortunate truth is that the data is not available, as our collection of target-based data was already far more limited than our phenotypic set. Further, it is no secret that merely collecting such data from available data sources can be misleading. Errors described above depict the challenges we faced in curation and collection; for example, a user looking for compounds active against NS2B-NS2 protease would not have found results due to the target being annotated generally as "genome polyprotein." We hope that our systematic analysis and enumeration of annotation deficiencies and bioactivity data curation protocols could help other researchers interested in expanding our collection or creating their own specialized collections. Most importantly, our efforts both identified several BSAs discovered by chance without deliberate focused efforts

(cf. Tables 1–2) as well as nominated several compounds for additional testing (cf. Table S9). As discussed above, the SMACC database included in this paper enables user-defined filtering of the data to support the generation of specialized subsets. In summary, we posit that this study provides strong motivation for continued investments into research targeting the discovery and development of novel BSA agents.

5. Conclusions

We have developed a pilot version of the SMACC (Small Molecule Antiviral Compound Collection) database containing over 32,500 entries for 13 emerging viruses. We identified eight compounds active against 3–4 viruses from the phenotypic data, 16 compounds active against two viruses from the target-based data, and 35 compounds active in at least one phenotypic and one target-based assay. Duplicates (phenotypic and overlap sets) and singletons (all sets) were also identified and annotated. While the pilot version of SMACC has integrated all chemogenomic data available in ChEMBL for these viruses, there was a large degree of sparsity (93%) within the integrated data matrix. Many viruses were understudied and thus, important results may be obtained by targeted testing of compounds included in SMACC against targets other than those they were tested against. In fact, we have suggested several such targeted testing experiments in this paper (cf. Table S9).

Our analysis indicates that not many BSAs have emerged from previous disconcerted studies and that special, focused efforts must be established going forward. The SMACC database built in this study may serve as a reference for virologists and medicinal chemists working to discover and develop BSA agents in preparation for future viral outbreaks. The SMACC database is publicly available in the form of a searchable Excel spreadsheet at https://smacc.mml.unc.edu.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Authors from UNC-Chapel Hill were supported by National Institutes of Health (Grants U19AI171292 and R01GM140154). Authors from NIH acknowledge support from the Intramural Research Program of the National Center for Advancing Translational Sciences (NCATS), National Institutes of Health (NIH). HJM acknowledges the support from the American Foundation for Pharmaceutical Education's (AFPE) Predoctoral Fellowship.

Glossary

8.

Broad-spectrum antiviral agent
an antiviral drug that shows activity against multiple viruses within, or between viral families

Emerging virus
a virus that holds a high impact for an endemic or pandemic viral outbreak event

9. References

Alves VM, Bobrowski T, Melo-Filho CC, Korn D, Auerbach S, Schmitt C, Muratov EN, Tropsha A, 2021. QSAR Modeling of SARS-CoV Mpro Inhibitors Identifies Sufugolix, Cenicriviroc, Proglumetacin, and other Drugs as Candidates for Repurposing against SARS-CoV-2. Mol Inform 40, 2000113. 10.1002/minf.202000113

Andersen PI, Krpina K, Ianevski A, Shtaida N, Jo E, Yang J, Koit S, Tenson T, Hukkanen V, Anthonsen MW, Bjoras M, Evander M, Windisch MP, Zusinaite E, Kainov DE, 2019. Novel Antiviral Activities of Obatoclax, Emetine, Niclosamide, Brequinar, and Homoharringtonine. Viruses 2019, Vol. 11, Page 964 11, 964. 10.3390/V11100964

Bobrowski T, Chen L, Eastman RT, Itkin Z, Shinn P, Chen CZ, Guo H, Zheng W, Michael S, Simeonov A, Hall MD, Zakharov AV, Muratov EN, 2021. Synergistic and Antagonistic Drug Combinations against SARS-CoV-2. Molecular Therapy 29, 873–885. 10.1016/j.ymthe.2020.12.016 [PubMed: 33333292]

Bobrowski T, Melo-Filho CC, Korn D, Alves VM, Popov KI, Auerbach S, Schmitt C, Moorman NJ, Muratov EN, Tropsha A, 2020. Learning from history: do not flatten the curve of antiviral research! Drug Discov Today 00, 1–10. 10.1016/j.drudis.2020.07.008

ChEMBL Compound Curation Pipeline [WWW Document], 2020. URL http://chembl.blogspot.com/2020/02/chembl-compound-curation-pipeline.html (accessed 7.7.22).

CRISIS2: A Phase 2 Study of the Safety and Antiviral Activity of Brequinar in Non-hospitalized Pts With COVID-19 - Study Results - ClinicalTrials.gov [WWW Document], 2022. URL https://www.clinicaltrials.gov/ct2/show/results/NCT04575038?view=results (accessed 7.7.22).

Erik DC, Guangdi L, 2016. Approved Antiviral Drugs over the Past 50 Years. Clin Microbiol Rev 29, 695–747. 10.1128/CMR.00102-15 [PubMed: 27281742]

Fourches D, Muratov E, Tropsha A, 2016. Trust, but Verify II: A Practical Guide to Chemogenomics Data Curation. J Chem Inf Model 56, 1243–1252. 10.1021/acs.jcim.6b00129 [PubMed: 27280890]

Fourches D, Muratov E, Tropsha A, 2015. Curation of chemogenomics data. Nat Chem Biol 11, 535–535. 10.1038/nchembio.1881 [PubMed: 26196763]

Fourches D, Muratov E, Tropsha A, 2010. Trust, but verify: on the importance of chemical structure curation in cheminformatics and QSAR modeling research. J Chem Inf Model 50, 1189–204. 10.1021/ci100176x [PubMed: 20572635]

Fu H, Zhang Z, Dai Y, Liu S, Fu E, 2020. Brequinar inhibits enterovirus replication by targeting biosynthesis pathway of pyrimidines. Am J Transl Res 12, 8247. [PubMed: 33437396]

Gaulton A, Hersey A, Nowotka M, Bento AP, Chambers J, Mendez D, Mutowo P, Atkinson F, Bellis LJ, Cibrián-Uhalte E, Davies M, Dedman N, Karlsson A, Magariños MP, Overington JP, Papadatos G, Smit I, Leach AR, 2017. The ChEMBL database in 2017. Nucleic Acids Res 45, D945–D954. 10.1093/nar/gkw1074 [PubMed: 27899562]

Li S.fang, Gong M. jiao, Sun Y. feng, Shao J. jun, Zhang Y. guang, Chang H. yun, 2019. Antiviral activity of brequinar against foot-and-mouth disease virus infection in vitro and in vivo. Biomedicine & Pharmacotherapy 116, 108982. 10.1016/j.biopha.2019.108982

Luthra P, Naidoo J, Pietzsch CA, De S, Khadka S, Anantpadma M, Williams CG, Edwards MR, Davey RA, Bukreyev A, Ready JM, Basler CF, 2018. Inhibiting pyrimidine biosynthesis impairs Ebola virus replication through depletion of nucleoside pools and activation of innate immune responses. Antiviral Res 158, 288–302. 10.1016/j.antiviral.2018.08.012 [PubMed: 30144461]

Melo-Filho CC, Bobrowski T, Martin H-J, Sessions Z, Popov KI, Moorman NJ, Baric RS, Muratov EN, Tropsha A, 2022. Conserved coronavirus proteins as targets of broad-spectrum antivirals. Antiviral Res 204, 105360. 10.1016/j.antiviral.2022.105360

Papadatos G, Gaulton A, Hersey A, Overington JP, 2015. Activity, assay and target data curation and quality in the ChEMBL database. J Comput Aided Mol Des 29, 885–896. 10.1007/s10822-015-9860-5 [PubMed: 26201396]

Park J-G, Ávila-Pérez G, Nogales A, Blanco-Lobo P, de la Torre JC, Martínez-Sobrido L, 2020. Identification and Characterization of Novel Compounds with Broad-Spectrum Antiviral Activity against Influenza A and B Viruses. J Virol 94. 10.1128/JVI.02149-19

Schultz DC, Johnson RM, Ayyanathan K, Miller J, Whig K, Kamalia B, Dittmar M, Weston S, Hammond HL, Dillen C, Ardanuy J, Taylor L, Lee JS, Li M, Lee E, Shoffler C, Petucci C, Constant S, Ferrer M, Thaiss CA, Frieman MB, Cherry S, 2022. Pyrimidine inhibitors synergize with nucleoside analogues to block SARS-CoV-2. Nature 2022 604:7904 604, 134–140. 10.1038/s41586-022-04482-x [PubMed: 35130559]

Sessions Z, Bobrowski T, Martin H-J, Beasley J-MT, Kothari A, Phares T, Li M, Alves VM, Scotti MT, Moorman NJ, Baric R, Tropsha A-D, Muratov EN, 2022. Praemonitus praemunitus: can we forecast and prepare for future viral disease outbreaks? Authorea Preprints. 10.22541/AU.166490960.04750244/V1

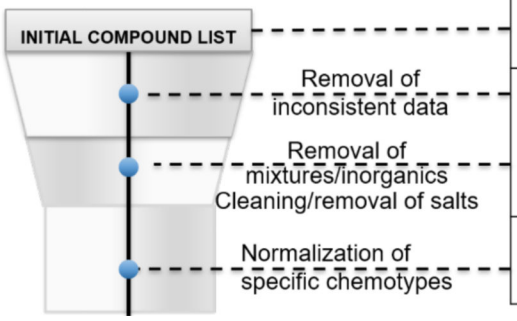
The BioAssay Ontology (BAO) [WWW Document], 2022. URL http://bioassayontology.org/(accessed 7.7.22).

Wang Q-Y, Bushell S, Qing M, Xu HY, Bonavia A, Nunes S, Zhou J, Poh MK, Florez de Sessions P, Niyomrattanakit P, Dong H, Hoffmaster K, Goh A, Nilar S, Schul W, Jones S, Kramer L, Compton T, Shi P-Y, 2011. Inhibition of Dengue Virus through Suppression of Host Pyrimidine Biosynthesis. J Virol 85, 6548–6556. 10.1128/JVI.02510-10 [PubMed: 21507975]

Yazdani S, Maio N. de, Ding Y, Shahani V, Goldman N, Schapira M, 2021. Genetic Variability of the SARS-CoV-2 Pocketome. J Proteome Res 20, 4215. 10.1021/acs.jproteome.1c00206

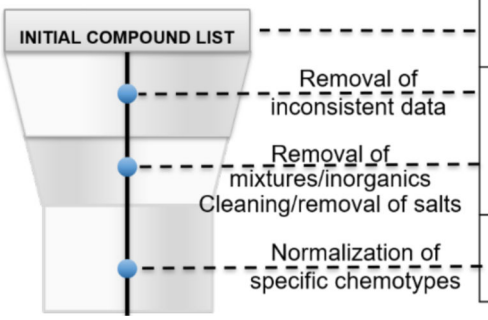
Highlights:

- We created a curated Small Molecule Antiviral Compound Collection (SMACC) database.
- SMACC covers 13 emerging viruses from 5 families.
- SMACC is available online at https://smacc.mml.unc.edu.
- We have identified ~50 compounds with broad-spectrum antiviral activity.
- Broad-spectrum compounds mostly cover flavi- and coronaviruses.



Coronaviridae	*Orthomyxoviridae*	*Paramyxoviridae*	*Phenuiviridae*	*Flaviviridae*	**Total**
SARS-CoV-2, MERS-CoV, HCoV-229E	H1N2 H7N7	RSV HPIV-3	Sandfly	Dengue Zika Yellow Fever Powassan West Nile	
18,190	**81**	**4,363**	**24**	**4,492**	**27,150**
13,389	74	4,013	18	3,898	21,392
13,389	**74**	**4,013**	**18**	**3,898**	**21,392**

Figure 1.
The effect of phenotypic assay data curation on reducing the resulting dataset sizes.



Coronaviridae	*Orthomyxoviridae*	*Paramyxoviridae*	*Hantaviridae*	*Flaviviridae*	Total
SARS-CoV-2, MERS-CoV, HCoV-229E	H7N7	RSV HPIV-3	Sin Nombre	Dengue Zika West Nile	
8,883	**26**	**150**	**130**	**1,949**	**11,138**
8,870	26	150	130	1,947	11,123
8,870	**26**	**150**	**130**	**1,947**	**11,123**

Figure 2.
The effect of target-based assay data curation on reducing the resulting dataset sizes.

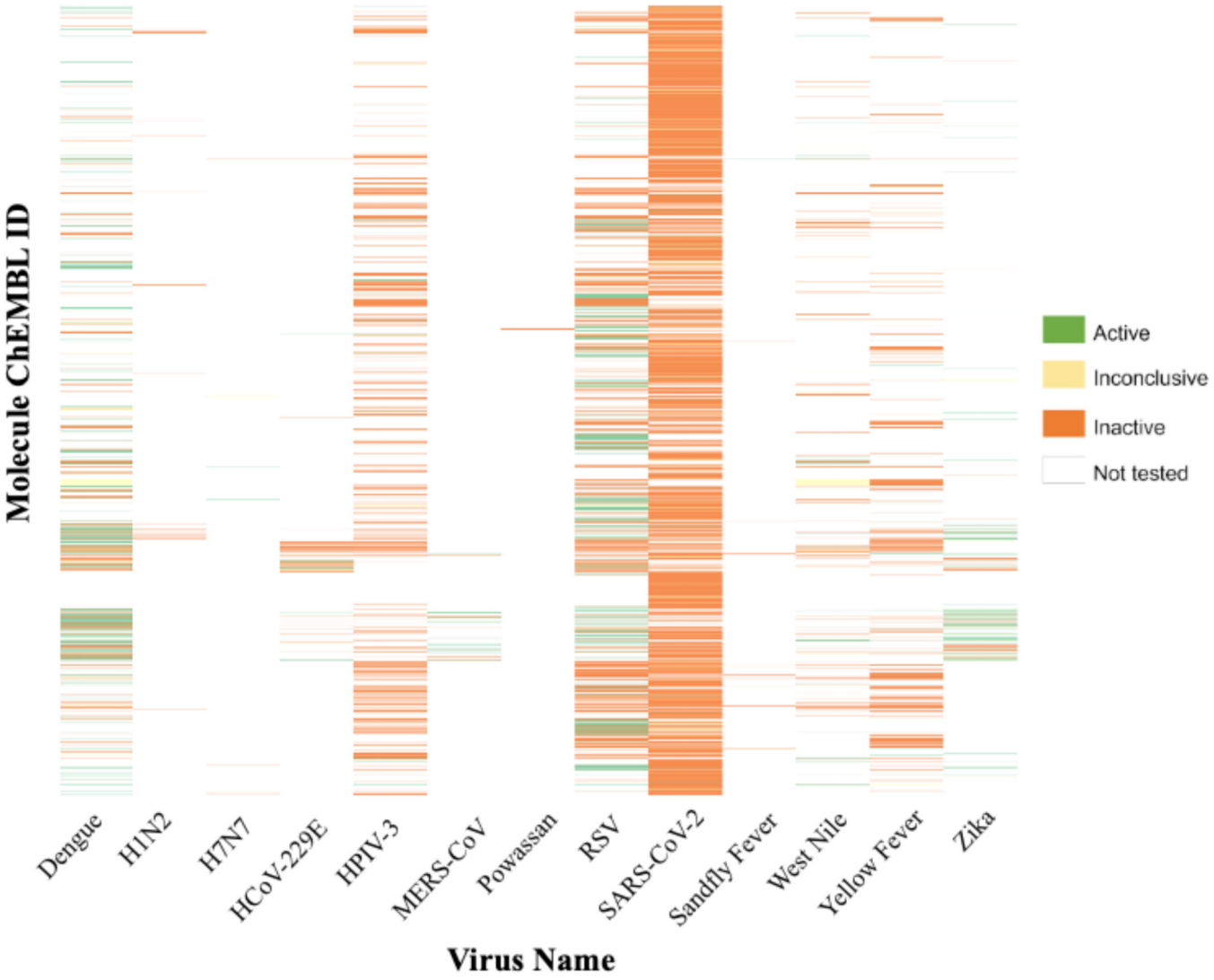


Figure 3.

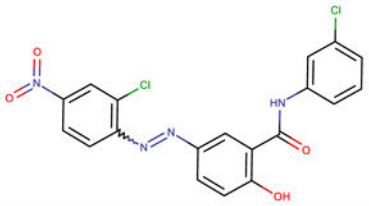
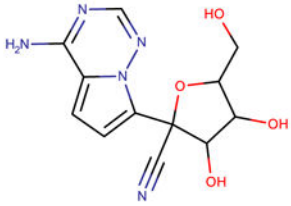
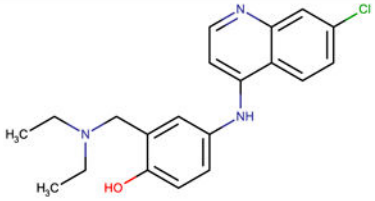
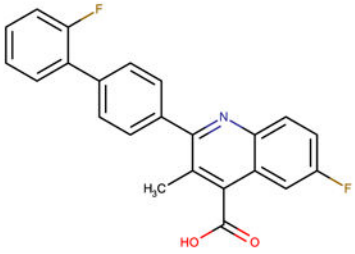
Activity heat map for 21,392 compounds tested in phenotypic assays for the 13 viruses selected for the SMACC database.

Figure 4.

Activity heat map for 11,123 compounds tested in target-based assays for the 10 viruses selected for the SMACC database.

Table 1.

Example of compounds active in multiple viruses.

ChEMBL ID	Structure	Active against	Inactive against
CHEMBL4437334		Dengue, West Nile, Yellow Fever, Zika	
CHEMBL4454780		RSV, MERS-CoV, Dengue, Zika	
CHEMBL2016757		RSV, HPIV-3, Dengue	West Nile
CHEMBL564201 (6-Azauridine)		RSV, West Nile, Dengue	Yellow Fever
CHEMBL682 (Amodiaquine)		SARS-CoV-2, Dengue, Zika	
CHEMBL38434 (Brequinar)		Dengue, West Nile, Yellow Fever	SARS-CoV-2

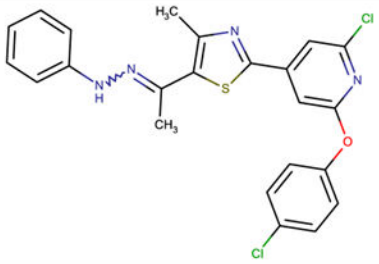
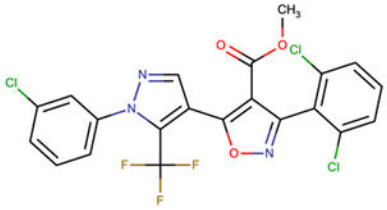
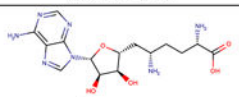
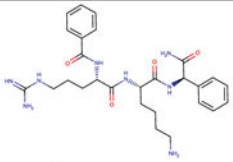
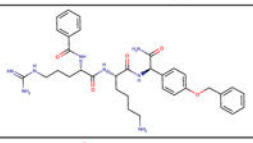
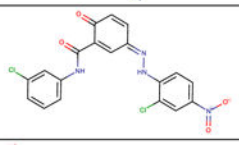
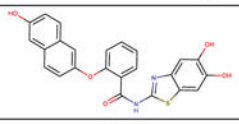
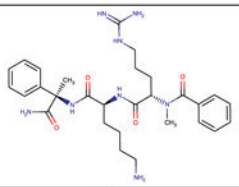
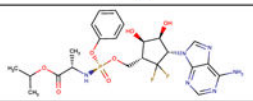
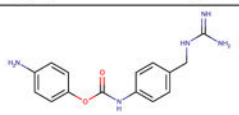
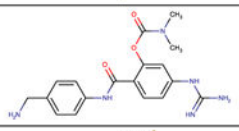
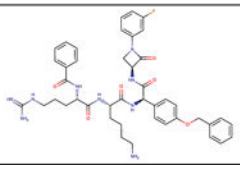
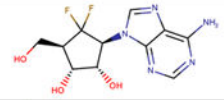
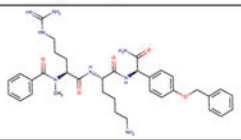
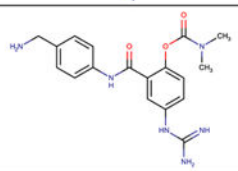
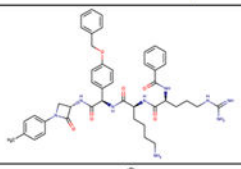
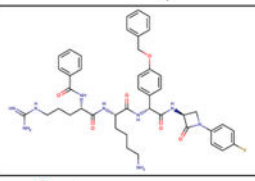
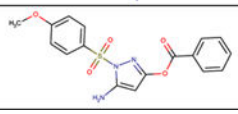
CHEMBL4544911		Dengue, West Nile, Yellow Fever
CHEMBL4562509		Dengue, West Nile, Yellow Fever

Table 2.

Compounds active against different viruses in target-based assays.

Compound Name	Structure	Target	Virus
CHEMBL1214186[a]		NS5	Dengue, Zika
CHEMBL3740277		NS2B-NS3 Protease	Dengue, West Nile
CHEMBL3741422		NS2B-NS3 Protease	Dengue, West Nile
CHEMBL4437334		NS2B-NS3 Protease	Dengue, Zika
CHEMBL4440832		NS2B-NS3 Protease	Dengue, Zika
CHEMBL4474101		NS2B-NS3 Protease	Dengue, West Nile
CHEMBL4522602		NS5, RDRP	Zika, MERS-CoV
CHEMBL4531546		NS2B-NS3 Protease	Dengue, West Nile
CHEMBL4536920		NS2B-NS3 Protease	Dengue, West Nile
CHEMBL4537775		NS5, RDRP	Zika, MERS-CoV

CHEMBL4544781		NS2B-NS3 Protease	Dengue, West Nile
CHEMBL4545026		NS2B-NS3 Protease	Dengue, West Nile
CHEMBL4563372		NS2B-NS3 Protease	Dengue, West Nile
CHEMBL4568434		NS2B-NS3 Protease	Dengue, West Nile
CHEMBL4576745		NS2B-NS3 Protease	Dengue, West Nile
CHEMBL569561		NS2B-NS3 Protease	Zika, West Nile

[a] Inactive against SARS-CoV-2

Table 3.

Approved drugs with active assay results found in SMACC.

Compound Information					Active Assay Results in SMACC	
Drug name	**Brand name**	**Approved clinical use**	**Inhibitory MOA**	**ChEMBL ID**	**Phenotypic Assay**	**Target-Based Assay**
Simeprevir	Olysio®	HCV	NS3/NS4B Protease	CHEMBL501849		H7N7 Matrix Protein 2
Asunaprevir	Sunvepra®	HCV	Protease	CHEMBL2105735		H7N7 Matrix Protein 2
Sofosbuvir	Sovaldi®	HCV	NS5B	CHEMBL1259059	Dengue Virus	
Ribavirin	Copegus®	HCV, RSV, fever	RdRp	CHEMBL1643	HPIV-3, Sandfly Fever, Dengue Virus, Yellow Fever Virus, RSV	
Lopinavir	Kaletra®	HIV	Protease	CHEMBL729	SARS-CoV-2	
Nelfinavir	Viracept®	HIV	Protease	CHEMBL584	Dengue Virus	
Raltegravir	Isentress®	HIV	Integrase	CHEMBL254316		H7N7 Matrix Protein 2
Elvitegravir	Vitekta ®	HIV	Integrase	CHEMBL204656		SARS-CoV-2 3CLpro
Atazanavir	Reyataz®	HIV	Protease	CHEMBL1163		SARS-CoV-2 3CLpro
Rilpivirine	Edurant®	HIV-1	Nonnucleoside reverse transcriptase	CHEMBL175691	SARS-CoV-2	
Podofilox	Condylox®	HPV-related diseases	Cytotoxicity/cell division	CHEMBL61	SARS-CoV-2	
Trifluridine	Viroptic®	HSV	Viral and cellular DNA synthesis	CHEMBL1129	HPIV-3	
Idoxuridine	Dendrid®	HSV-1	Viral and cellular DNA synthesis	CHEMBL788		Zika NS5
Acyclovir	Zovirax ®	HSV, VZV	Viral DNA polymerase	CHEMBL184		H7N7 Neuraminidase
Zanamivir	Relenza®	Influenza A and B	Neuraminidase	CHEMBL222813	H7N7	